TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

KHURSHID ALI QURESHI

DATA COLLECTION AND INFORMATION FLOW MANAGEMENT FRAMEWORK FOR INDUSTRIAL SYSTEMS

Master of Science thesis

Examiners: Professor Dr. Jose L. Martinez Lastra, Dr. Andrei Lobov
Examiner and topic approved by the Council meeting of the Faculty of Engineering Sciences on 5th October 2016

ABSTRACT

KHURSHID ALI QURESHI:
Tampere University of Technology
Master of Science Thesis, 71 pages
April 2017
Master's Degree Programme in Automation Engineering
Major: Factory Automation and Industrial Informatics
Examiner: Professor Dr. Jose L. Martinez Lastra, Dr. Andrei Lobov
Supervisor: Angelica Nieto Lee

In today's global era of competitive environment, the importance of data management is critically evaluated. From business decision making to inventory management, the information generated from data which is gathered from processes and systems is incredibly valued for optimization and analysis. Often the data is transferred and integrated to get the information needed for the big picture of the organization. This interoperability reluctance of the rigid legacy systems deployed in organizations is one the major hindrances of information sharing. This issue has been addressed in the thesis with facts and details and its comparatively reliable solution has been presented.

With the evolution of technology, the supply chain is able to categorize the large amount of data into information required by the user in cost effective and time efficient way. Although ERP systems have been a major breakthrough in streamlining the supply chain management (SCM), cloud computing has revolutionized the SCM. For instance, ERP produces a huge amount of data during the production processes, the real time visibility and access to these remote data sources has been made possible by cloud computing.

The aim of this research is to provide a user-friendly data collection framework through which reliable data can be acquired anytime according to the JSON format provided by the user. To achieve this objective, the research has been conducted in two parts: theoretical research and empirical research. In the first part, the detailed theoretical background of data management, legacy systems, information flow and function blocks have been analyzed. Whereas, the empirical research focuses on the cloud based computing platform, called cloud computing network (C2NET), and implementation of two use cases to resolve the data collection and information flow issue in the industrial legacy systems.

The result of this research work proposes a platform for unifying the flow of information from ERP systems with cloud based systems according to the requirements of the user, which in this case is C2NET platform. This key function is done by executing function block based approach supported by PlantCockpit which uses IEC 61499 standard and the service oriented architecture (SOA) project results. The data or messages received from heterogeneous legacy systems are synchronized in the Legacy System Hub and transferred to the target system with the help of REST, SQL and XML adapters. In this way, the integration of legacy system is carried out along with the harmonization of acquired data.

This unification of data from legacy system through cloud computing network has made the efficient and timely collection of accurate and reliable data easier for the user. It allows the user to extract the information from industrial systems readily available according the needs and thus, is able to mitigate the interoperability issue of legacy systems.

PREFACE

After relentless efforts and continuous struggle for several months, I was able to accomplish this research work. From shortlisting the title to the planning and execution of the thesis, I invested a great amount of time in trying to produce a well-researched and thoughtful manuscript.

I am proud to say that in retrospect, the project gave me valuable insights about the research area I loved to explore. During the research, I could foresee myself drowning into the ocean of exciting learning. Consequently, after the thesis, without any hesitation, I can vouch that I have been able to gain enough knowledge in this research area and I want to explore it more in future.

Fortunately, I have been gratified with the support and guidance of lot of people. Foremost, I would like to express my sincerest gratitude to Dr. Andrei whose engagement, feedback and guidance throughout this learning phase steered me in the right direction and enabled me to complete my thesis. Further, I would like to thank my supervisor Angelica, who skillfully supervised my work on each step. Their exceptional supervision and guidance kept me motivated and enthusiastic about this work piece.

Regarding the specification and technicalities of the thesis, I would like to thank my incredible colleagues Borja and Wael. Their fruitful cooperation has been unmatchable and I am blessed to work with them. In addition to that, I am highly grateful to my childhood fast friends Ateeb and Huzaifa who provided immense guidance in my thesis through their expertise and willingly shared their precious time during my research. In addition, I would like to thank my friends Arsalan and Adnan for their technical support while carrying out this thesis and Ahsan for helping me finalize the writing process.

I would like to express my gratitude to the entire FAST Lab who provided me amazing work environment. The learning opportunity and friendly platform they gave me has been a major success factor in completing my thesis. I gratefully acknowledge Prof. Dr. Jose L. Martinez Lastra and Anne Korhonen for all the support and encouragement during the course of the research.

Last but not the least; I would like to thank my wonderful family and friends for always being there for me. Especial thanks to my beloved parents who are my biggest motivation to accomplish my dreams and my better half, my wife for walking me through every single step, motivating me, helping me, listening to me even when she didn't had the slightest clue what I was talking about and giving me the strength to finish this thesis in its entirety.

Thank you all from the core of my heart!

Khurshid Ali Qureshi

4th April 2017

Tampere, Finland

CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

C2NET	Cloud Collaborative Manufacturing Networks
DCF	Data Collection Framework
DCS	Distributed Control System
ERP	Enterprise Resource Planning
ESB	Enterprise Service Bus
FB	Function Block
FBI	Function Block Instance
FBM	Function Block Manager
FTP	File Transfer Protocol
HTML	Hyper Text Markup Language
HTTP	Hypertext Transfer Protocol
ICT	Information Communication Technology
IEC	International Electrotechnical Commission
ISA	International Society of Automation
JBI	Java Business Integration
JMS	Java Messaging Service
JSON	JavaScript Object Notation
LSH	Legacy System Hub
MES	Manufacturing Execution System
OSGi	Open Service Gateway
PCP	PlantCockpit
REST	Representational State Transfer
RM	Resource Manager
SAP	System Application Product and Products
SCM	Supply Chain Management
SDA	Simple Data Adapter
SFTP	SSH File Transfer Protocol
SM	Source Manager
SQL	Structured Query Language
TUT	Tampere University of Technology
URL	Uniform Resource Locator
XML	Extensible Markup Language

1. INTRODUCTION

The first chapter of this thesis will give an overview of the problem statement. In the light of the background given to introduce the topic, the objective and motivation of this thesis is expressed in this section. It will also emphasize briefly on the scope of the thesis to engage and enlighten the target audience.

1.1 Motivation

With the globalization of today's economy, the time efficient way to manage the supply chain is essentially required. The aging of systems is necessary to be accommodated promptly so that they can be revitalized according to the current technology. Otherwise, the misalignments of operating systems can cause disruption of supply chain, which will result into halting the process with waste of time and money.

Legacy systems are those data sources, which are facing obsoleteness but are vital to the organizations operations on larger scale. Often software has some glitches and act abnormally. Apart from the torturous maintenance and upgradation of legacy systems, their execution of businesses processes is also rigid and has predefined process flows, which downgrades customer relationship management software. Organizations want to have some reliable and flexible systems that meet the requirements of not only that particular organization but also the needs of customers. Despite the challenges legacy systems present, they are hard to replace due to high cost and lack of reliable technological solution. This issue has been discussed for a long time in the past too but failed to present the easiest way to overcome its weaknesses and none has been able to aptly cater the current need of organizations. The seamless integration of information flow for the maintenance of the legacy systems is required. Data migration, reverse engineering and data integration are the few methods of maintaining or upgrading or in short modifying the legacy systems.

Besides the aforementioned problems of legacy systems, they are the source of generating low quality, erroneous data in the organizations. The data is an asset of any organization. Any kind of data, which is inconsistent, missing, inaccurate or irrelevant, can be a huge loss for the organization. It can halt the processes of the organization and will direct to costly and poor decisions. Thus, the need to develop an efficient way of acquiring flexible and reliable data from the legacy system is required. Considering this gap in light, the research done in this manuscript presents a highly likely solution to deal with the data management issue.

The cloud based platform of Cloud Collaborative Manufacturing Networks (C2NET) project, which works on the real time integration of legacy systems, supports the data collection and information flow from legacy system. It gives an all-inclusive approach for fetching data from legacy system by using function blocks. The legacy system hub is the platform based on cloud computing technology, which will gather data from various legacy systems and integrates that data into user-desired information format with the help of adapters.

1.2 Scope of Thesis

This thesis will shed light on a different and an efficient way of countering the challenges of legacy systems specially data collection and information process flow which have not been thoroughly tackled together earlier. The solution to accommodate primarily the data management issue of legacy system will be discussed in this thesis. Various outdated sources of data like: 'a table in an Structured Query Language (SQL) database, a simple text file, an Extensible Markup Language (XML) document, a spreadsheet, a Web service, a sequential file, a Hyper Text Markup Language (HTML) page, and so on' [16] are the legacy systems from which the data acquisition is difficult. In order to have a data integration strategy, an essential part of data collection framework will be redesigned is such a way that it will become easy to fetch data through the Legacy System Hub (an integration of different simple adapters, used as a medium of communication between legacy systems and data collection framework). This thesis aims to provide the functionality of C2NET platform. In addition to that, the role of SQL, Representational State Transfer (REST) and Excel adapters in fetching the data from Legacy System Hub (LSH) and C2NET platform through the PubSub module is the main agenda of this manuscript.

In this way, industrial setups will have an aggregate view on the entire network, which will allow them to have a swift and authentic feedback system to get the required information gathered through collection of data from heterogeneous data sources. It will in turn enable the companies to promptly react to market changes and become more productive.

1.3 Hypothesis

The acquisition of data from the legacy system data sources like Excel Sheets, REST Endpoints and SQL Databases will be made time efficient and reliable by using the major components (Legacy System Hub, Resource Manager and PubSub) of the designed data collection framework (C2NET).

1.4 Objective

The overall objective of this study is to design a data integration strategy in such a way, that it will allow the user to easily access authentic data by collecting it from different data sources and then converting it into JSON format as per user requirement.

To meet this objective this thesis will focus on the following:

1. Redesigning the data collection framework
2. Giving an easier way to fetch data for convenience of the user by providing a single platform (a hub) to access all the information from different, multiple data sources.
3. Returning the data in a structure provided by the user
4. Providing fast knowledge feedback loop to companies. The data fetched is readily available for the user to check for errors through an interface

1.5 Structure of Thesis

Broadly, this thesis is divided into two main sections, which are theoretical research and empirical research. However, for the detailed analysis of each step of the research conducted, the thesis consists of total six chapters. The description of which are as follow:

Chapter 1 defines the problem statement, the motivation to study the scope of research, the hypothesis generated to achieve the objective and the structure of overall thesis.

In Chapter 2 the detailed analysis of the previous research has been elaborated. This chapter forms the basis of research and enlightens the readers about the concepts involved in the scope of the study. It is the most extensive part of the thesis as it deals with all the relevant theories and stresses the problem in reference to these theories or research work done in the past.

Chapter 3 describes the research methodology adopted to achieve the set goal. It involves the research questions designed to realize the problem statement. This chapter also deals with the method of research and approach used to find the solution of the problem.

After laying the foundation of the theoretical background and research methodology, Chapter 4 introduces the empirical research done. It discusses the implementation of the data collection framework developed to meet the set objective.

Chapter 5 forms the discussion about the results. It covers the limitations of the thesis and revisits the problem and research questions to analyze the results. This chapter carries a huge significance in investigating the success and failures of the theoretical and practical research of the thesis.

Finally, Chapter 6 concludes with the summary of the results and recommends areas for future research work. It also encompasses the significance or credibility of the thesis under the validation of research part.

1.6 Research Limitations

Although the research done in this thesis has been able to meet the set objective, the following were some methodological limitations, which were not avoidable due to time and content constraint.

Firstly, the collection of data is ensured from only three data sources that are REST, SQL and Excel. However, there are plenty of other resources of data generation, which were not taken into account in this thesis.

Secondly, SQL is used for fetching the data from only one type of SQL database that is MySQL, in this research. The ability of the designed adapters to work with other databases was considered in the initial implementation phase but was dropped later on due to the then created requirements of the C2NET project.

Thirdly, the Resource Manager (RM) interface for communication of messages with Legacy System Hub (LSH) is not generic and it is used only for these three use cases. That is why; to fetch the data from other databases, this approach of using RM cannot be executed in its current form.

Lastly, no large-scale usability studies were conducted to evaluate the efficiency of LSH compared to legacy systems. As a result, this thesis does not contain the quantitative benefits of LSH. These include the reduction in data acquisition time, the decrease in the number of errors, and the impact on the budget spent on maintenance of legacy systems.

2. LITERATURE REVIEW

This chapter will give an elaborated view of the past work done on the research area been examined. It will give a detailed theoretical framework of the research done so far on Data Management, Legacy Systems, Information flow, Industrial systems and Function Blocks. In addition, it will enable the reader to get the gist of the founding concepts of this thesis in order to answer the research questions.

2.1 Data Management

Data management is a process through which data is administered for users in such a way that it has to be reliable, secure and error free. Data Management International (DAMA) gives the most appropriate definition of data management: [78]

"Data Resource Management is the development and execution of architectures, policies, practices and procedures that properly manage the full data lifecycle needs of an enterprise."

Similarly, another definition of data management is provided by DAMA Data Management Body of Knowledge: [78]

"Data management is the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets."

The collection of authentic data is the most vital part of any support system. The data collected manually is often faulty, outdated, misrepresented and inaccurate [18, 19, 20, 21, 22]. McCullouch points out that manual collection and recording of data used to take 30 to 50 percent of field supervisors' time [19]. The usage of data is one of the major day-to-day tasks of organizations. Therefore, an efficient way to acquire high quality of data is essential for the companies to succeed [8]. To better understand the concept of data quality, we can refer it as "fitness for use'' as it focuses on user convenience [9, 10, 11, 12]. The most frequently indicated data quality evaluation measures are consistency, accuracy, timeliness and completeness [1, 7]. These measures have been elaborated in Figure 1 below. In [13], Marsh lays the importance of data as the 'most valuable asset of any' business. For him, if the data somehow fails to achieve any of these measures, it would lead to useless and costly decisions for any organization. He states that businesses are not themselves aware of the bitter reality that they are suffering from low quality data unless some legacy system stops working due to loss of source code or maybe due to the retirement of the person who wrote that code for legacy system [13].

Figure 1. *Data Management Measures*

Apart from this, a lot of research has been done on the consequences of poor quality data: ranging from economic problems (high operational cost) to negative social and cultural impact (data reliability issues and decreased job satisfaction) [2 ,4, 5, 6]. Despite of the stated problems of data quality and its detrimental effects on businesses, unfortunately it has not been worked upon intensively mainly due to lack of efficient technology driven solutions to acquire the data [13].

The importance of quality of data can be analyzed from the example, which Kanaracus gave in his article. He draws the attention that poor data quality led to complete system failure when a National Grid in New York implemented new SAP payroll system. Out of the several reasons like exaggeratedly ambitious design and improper training of employees, one of the major reasons was the poor quality of data in the legacy system [3].

2.1.1 Difference between Data and Information

It is essential to differentiate between data and information in order to delve deeper on the importance of data. Redman identifies the connection; data is a source of information, from which knowledge is derived and knowledge leads to wisdom [60]. Hence, data is the prime source of information, wisdom and knowledge, as mentioned in Figure 2.

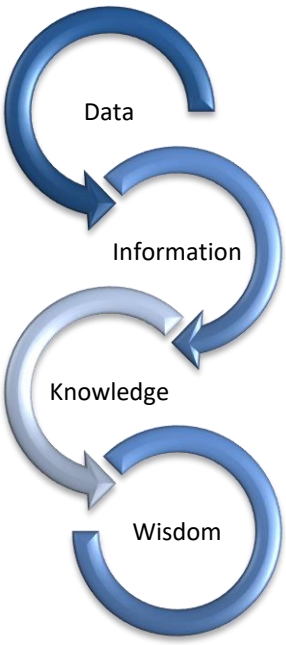


Figure 2. *Relationship chain of data information, knowledge and wisdom*

Data constitutes of the raw material for Information Age. However, this raw material has the tendency to be used repeatedly, contrasting to physical raw material [9]. Some authors define data as the source of providing the incoherent information. Similarly, data in terms of organizations perspective is defined as the basis of recording the transactions or day-to-day activities in a proper format [61]. Haug defines the data as the foundation of recording the observations and symbols about the significant event [64]. In contrast, Zack defines data as collection of irrelevant information of facts and observations [62].

Both Hague and Redman agree on the fact that data is essentially made up of two things. One is data model and other is its attributes. Data model consists of an entity like employee and its values are its attributes. Values can range from employee ID to his/her personal record (name, address, date of birth, etc.), whereas, data model is the major description. For instance, the data model is any 'employee' who has an attribute 'ID' the value of which could be '5671'.Thus; the data retained in this form becomes the record [63, 60].

Unlike other sources, data has multiple advantages; it can be copied, stored and accessed whenever required. These benefits come with challenges. For instance, data quality has to be ensured for proper use of data [63]. Similarly, Redman identifies some exceptional qualities of data as compared to other resources [60]:

• Data can be multiplied.

• Data can be shared, stored and combined.

• Data is organic.

• Data becomes the lingua franca of the organization.

• Data can be lost and retrieved.

• Data is the organization's way of keeping the information safe.

• Each organization has its own data.

• Data is not factual.

Some authors define information as data, which is gathered during the process of transformation. For them, information is a data, which is obtained from contextualization, calculation, concentration and specification of data [61]. Whereas, authors like Hague defines information as the data, which is used for particular purpose and is presented in a specific way [64]. However, Zack comes up with a unique perspective of defining the relationship between data and information. According to him, information is the data, which makes sense when placed in the form of text or have some meaning in some kind of message [62].

The Journey from Data to Knowledge

Data has no meaning unless it is understood in the relevance of context. As information is a series of messages, it is required that data should be transformed into information for proper interpretation. Deretske defines information as the flow of meaningful messages, which give us knowledge [85]. On the other hand, it is hard to define knowledge, as it is a multifaceted concept. However, it is about being explicit and tacit [86, 87] that is to 'know what' and to 'know how'.

The collection of data cannot be considered as information. Likewise, the collection of information is not knowledge. Hence, it is not about the collection; rather, it is about the synergy between these resources.

Knowledge Management and Information Systems

Knowledge management is about the procedure of acquiring the knowledge and the process through which it is obtained. It can consist of tacit or explicit framework for human understanding through which the belief becomes truth. The role of technology in supplementing the knowledge management cannot be denied. Thus, the importance of intranet and internet, which together consist of information system, has immense importance in retrieving the knowledge. The knowledge, which is represented, must be structured and information systems are vital for restructuring that. It means that the information is stored (data) in the system in a structured way so that answers can be found or in other words, knowledge can be derived from it easily [88].

A good information system is capable of the following characteristics:

- **User Friendly**
  Easily comprehends the query and asks for additional information if required but gives the response by thorough interpretation.

- **Flexibility**
  Readily adapts itself according to circumstances requested by user to give the information, which is required.

- **Easy Communication**
  It gives answers in easily understood layout for the convenience of user.

- **Reasoning**
  A good information system gives reasoning for its interpretation of the problem.

- **Facilitation**
  It facilitates the user in knowledge gathering through a structured and efficient process and framework.

- **Overarching**
  It has various techniques, structures and formats for delivering and acquiring the data stored in it.

- **Resourceful Tools**
  An expert information system comprises of numerous resourceful tools for language interpretations, graphical demonstration and information generation, which helps in the suitable representation of knowledge.

Transformational Process: Data, Information, Knowledge and Wisdom

It is essential to understand data, information, knowledge and wisdom in relation to context. Context is something, which gives analysis, identification and meaning to these resources. It gives validation to these resources. In order to study the transformational process of data, information, knowledge and wisdom, context is used.

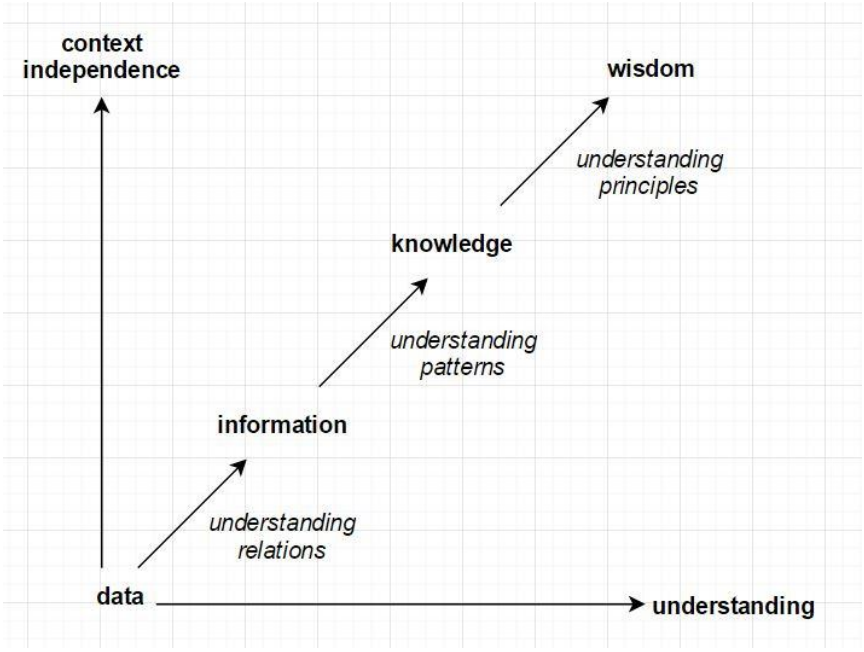


Figure 3. *Relationship between Context and Understanding; Explanation of Transformational Process [89]*

From the above figure, it is inferred that information just gives the understanding of relation between data. It does not tell the reasons of process, the why, how and what of data. Information is a linear and static resource. Thus, information greatly depends on the context to explain the related meaning. Whereas, knowledge gives the understanding of patterns, the process and the reasons, it includes the strategy and analysis of how the things are done [88].

2.1.2 Quality of Data

So far, the meaning and importance of data has been discussed. Now it is essential to figure out the qualification of data. Data quality is the most essential feature of adequate data. It is not easy to define, thus, need some in depth analysis of what can be inferred from quality and what measures can be adopted to check the quality of data [1]. Hence, the data quality is discussed in detail below.

As discussed above as well that data refers to "*fitness for use"*, it implies that quality of data is adequate if it is able to serve the perceived purpose [9, 63]. Thus, data quality becomes meaningful if it is placed in some context [9]. For Haug & Arlbjørn the quality of data is a relative concept. Further, Tayi and Ballou conclude that data quality, which is useful for one purpose, cannot necessarily be suitable for other purposes [9].

The quality of data is a concept encompassing various dimensions [2]. Hence, it is a multidimensional concept [7, 1, 65]. There are different parameters to measure the quality and to manage the data efficiently [66]. Thus, having different dimensions to figure out the quality of data is another way to get to know the data quality.

As mentioned earlier there are four most frequently used data quality measures. Apart from these, there are various other dimensions defined by different authors [67, 66]. Nonetheless, it is hard to acquire the data, which is not subject to error. In addition, it is not a requirement to have error free data. The data is just supposed to fulfill the criteria of user. Thus, the quality of data varies from its usage and applications [68].

In summation, it is concluded that the quality of data is a kind of complex concept. However, it is essential to get the hang of it, in order to measure and manage the data effectively. For this reason, there are various measures to check the quality of data. Moreover, the ability to measure the quality of data can be enhanced by considering the various dimensions of data quality. [66]

2.1.3 Concerns of Data Quality

Data quality is an intricate process. It is difficult to measure the quality of data considering the various aspects involved in it. In addition, while measuring the quality of data, it is hard to take

account of all the possible conditions, which can affect the data quality [9]. Hence, it is important to be apprehensive of all the possible problems data can have.

Redman has identified the following seven problems of data quality in his article [60]:

1. Unorganized data
2. Data reliability and security
3. Difficult to find relevant data
4. Erroneous data
5. Wrong interpretation of data
6. Data recognition and description problem
7. Confusion within organizational data

Redman further builds on this point that the studies conducted by well-known authors have deduced that almost 30 percent employees waste their productive time in search of non-erroneous data which they require [60]. This indicates the difficulty the employees went through in finding the relevant data in a time efficient way. The time spent on acquiring the perfect yet relevant data is the differentiating fact between the victors and defeaters. It is because the world has entered into global information age and the asset of any organization depends on its ability to manage data time effectively. Therefore, if it is difficult to find the relevant data efficiently, it can result into various grave consequences like loss of profit, wastage of time and efforts, improper decisions, etc. [70].

The following figure shows how Redman categorizes the data quality problems into three sections of operations, decision making and strategy for organizations:

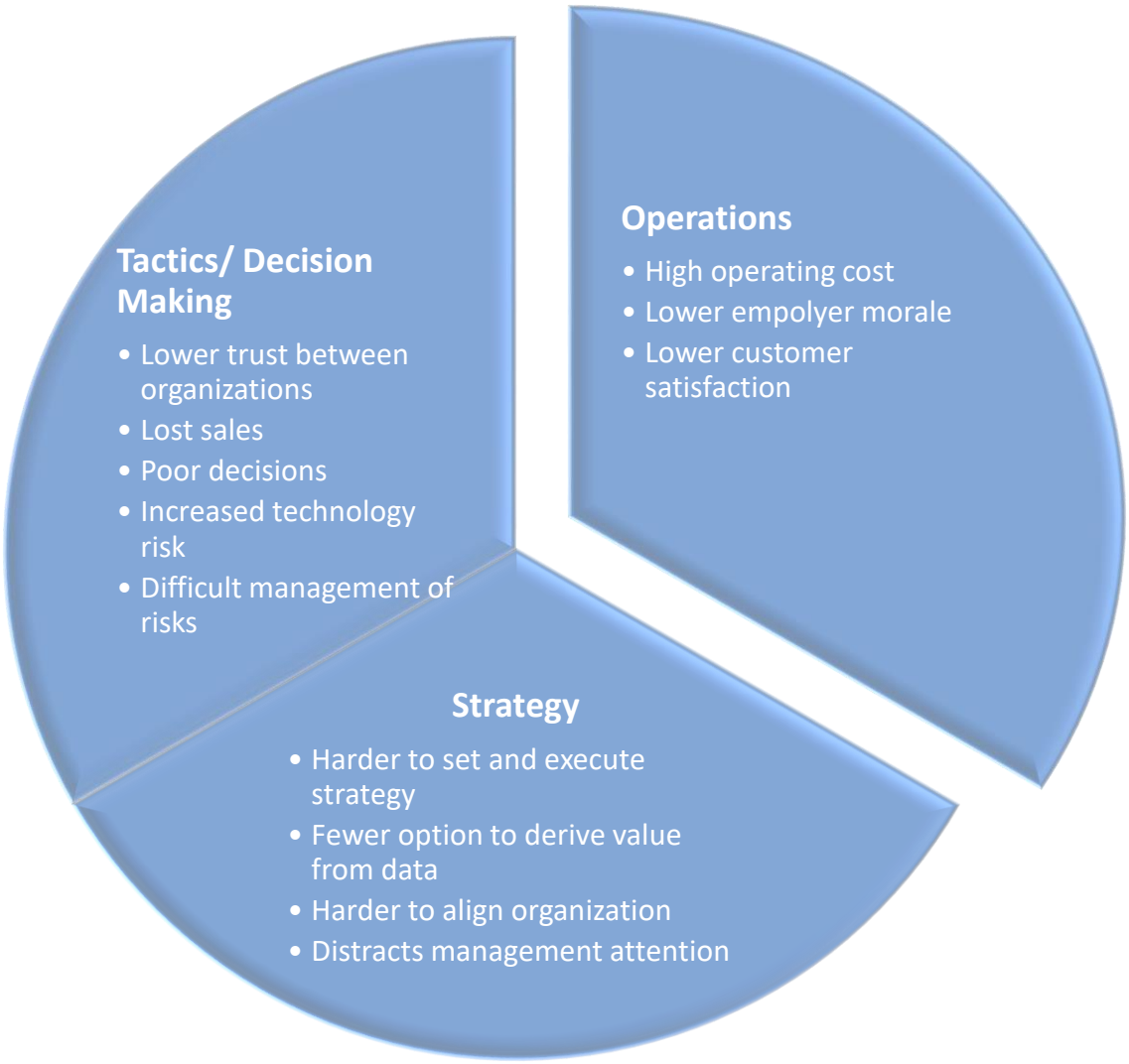


Figure 4. *The consequences of data quality problems by Redman [60]*

The above Figure 4 gives an idea of data quality problems on a larger scale, which means, that being the asset of the organization, data quality gravely affects the three major divisions of any organization. Subsequently, the following theory will explain the most frequently reported reasons of having low quality data.

Missing Data

Inaccurate data causes a lot of trouble in understanding, as it is not according to the values, which occur in the real world. According to one estimate, there is 25 to 30 percent faulty data inclusive of missing data [60]. One fact should be taken into account while calculating the erroneous data that the databases vary from organization to organization. Apart from missing and incorrect data, there is another issue of irrelevant data, which can be stated as correct. For instance, the social security of

the wrong person is found but it is valid because it exists and belongs to another person [68]. Hence, all the factors should be catered while looking for proper data.

Expired Data

Another feature of data quality is timeliness. It means that the data should not be expired. The example, which Tayi and Ballou give to understand the concept of timeliness of data, is of foreign exchange data. They say that if the newspaper prints the foreign exchange rates in the midnight and if the paper is printed and distributed in the morning, even then the data is expired. As the foreign exchange market is so volatile that it changes within hours. Further, they talk about inconsistency and missing data problem. For instance, the data can be both correct and timely yet it can be inconsistent due to some missing information. [9]

Huge Amount of Data

Other than quality issues of data there are some quantity issues too which in turn affect the quality of data. For instance, every operation in the organization creates more data, every decision leads to further addition to data, even a small event, as petty as taking the order of client creates data. However, major part of this vast amount of data is never used and required. Thus, it becomes quite difficult to store and retrieve the relevant data at time of needs.

Inconsistent Data

Huge amount of data causes the inconsistency in data. It can lead to an issue where two similar data entries, which might change later with time, result into data discrepancies. As the undistinguishable data can be used by two different departments of the same organization, so it can create confusion among employees afterwards [60]. One of the reasons of this confusion is the division of an organization into different departments and sub units [68].

Insecure Data

Cybercrime is becoming more and more common with the evolving technology. The hackers have been looking for important data to manipulate it since a long time [60]. Therefore, organizations have to be careful about the privacy of crucial data. Even a slight negligence in security of data can lead to serious loss of most valuable resource of the organization.

Poor Data Definitions

Unclear and confusing definitions of data are also one of the features affecting the quality of data [60]. The definition and understanding of data should be consistent in all the departments of any organization. Otherwise, it will cause a lot of trouble. For instance, some departments define customer data and contain irrelevant information –this data, if used, would lead to problematic result for profit margin and sales too, due to variation in values [68].

Confusing Data

Though companies have vast amount of data, yet they are confused about its usage and sufficiency. They do not know which data is relevant, how to prioritize it, from where it can be acquired, etc.

This creates a lot of confusion about data within the organization. According to Redman if we try to work on the improvement of quality of that data which will never be used, it will cause loss of time and resources [60]. As mentioned earlier also that organizations are themselves not aware of the quality of data they have and are reluctant to make any changes into them [69].

Considering the importance of data in collecting the information and utilizing it in the form of knowledge, the problems of data quality have to address in some way or another. This has been done in detail in the section 2.1.2 of this thesis, which highlights the methods of modernization of legacy systems. Therefore, to find the source of these issues related to data quality, it is necessary to study the source of data generation. Hence, the next part of the thesis will delve deeper into the examination of the system that is responsible for source of faulty, inaccurate and inconsistent data. These complex systems are legacy system, and below mentioned part give their elaborated analysis.

2.2 Legacy Systems

Bennett defines legacy systems as *"large software systems that we don't know how to cope with but that are vital to our organization"* [22]. Likewise, they are also famous as complicated and complex systems, which have been existing for several decades and are difficult to replace [27, 28, 29, 30]. Both Benett and Sneed have defined the legacy system as an 'outdated technology'. Some authors have discussed about the oldness of legacy systems [32, 33 34 35], while others have mentioned their compatibility, complexity, agility and risk [36, 31, 37].

2.2.1 Issues of Legacy Systems

On one hand, legacy systems have the most critical role to play in an organization, as the whole businesses software is dependent upon them. On the other hand, they are expensive to maintain, have large source codes and are difficult to replace. With time, they are becoming obsolete and face the challenges of compatibility with modern technological systems. The time taken to fetch the data from these systems is long, managing the data is inefficient, the data attained from them is unreliable and the information and expertise required for running them is difficult to acquire. They are usually known as legacy systems because of their superseded way of managing and acquiring data. It is essential to replace them with efficient technological innovation in order to mitigate the challenges they present. Despite these challenges, legacy systems have the most important role to play in the organizations as they are still in businesses and provide the main support for information flow [43]. Following are some of the problems of legacy systems discussed in detail in Figure 5:

Figure 5. *Common Issues of Legacy Systems*

High Cost of Operation:

The cost of maintaining, operating and supervising the legacy system is too high. It is so expensive that a survey analysis found that almost 85 to 90 percent of the company's total budget is spent on the maintenance of legacy system [44]. This implies that the company is only left with a little budget to spend on other necessary activities. In short, the operational cost of the legacy system can get the company into major financial trouble.

Maintenance Problem:

The programming languages, which were used to design legacy system, are not up to date and the world is left with very few experts of it. Therefore, it is utterly difficult to update the legacy system, as its software is also outdated. Further, the codes of this system are also very complex due to the 'ad hoc evolution' of legacy system [49]. No work has been done on the documentation of the legacy systems and they have been maintained in emergencies only [37]. Hence, it is very hard to maintain and improve them [28, 27and 46].

Fewer Experts

The qualified people, who know how to deal with legacy system are either getting old or have died. Therefore, very few experts know how to operate them. In addition, the new generations are more interested into new technologies and do not want to study the obsolete methods. According to survey, also the number of students who enroll in computer sciences is also dropping. For instance, the percentage drop in USA is 39% [47].

Outdated hardware

One of the reasons of being slow of legacy systems is that their obsoleteness of hardware [46]. They are at a high risk of failure because both their hardware and software are unable to be updated. This

is because the suppliers of hardware are also less in the market and mostly have stopped supplying it.

Inadequate Structural Design of Legacy System

The structure of legacy system is messily designed. It is a massive structure without proper distinction between user interface and other models [35]. They have rigid structure, which is not compatible with other systems and is less open to any external hardware or software [43].

Absence of Proper Documentation

The legacy systems suffer from the problem of updated documentation. Firstly, because of lack of knowledge and secondly, due to retirement of experts dealing with the legacy systems [48]. It can be because they might be devoid of enough information or they might have some structural ambiguity, which caused the absence of documentation on paper [37]. Further, whenever the system faces some trouble, it gets hard to trace out the problem due to lack of documentation and proper know how of the structural design [46].

2.2.2 Modernization of Legacy Systems

Nowadays, enterprises require such systems, which are reliable, manageable and affordable. These along with few other parameters like flexibility and scalability are considered the selection criteria for some businesses. Thus, considering these factors, the aforementioned issues of legacy systems make their survival problematic. Due to the rigid approach of legacy system hardware and software, it becomes difficult to keep them in organizations. Hence, the numerous issues of legacy system have derived the attention towards modernization of legacy system in order to make them compatible with the current technologies.

As mentioned above due to the slow processing of legacy system, they respond sluggishly to rapid market changes. Therefore, it becomes mandatory to modernize the legacy system so that they can respond effectively to business needs and can keep up with the technological advancement [50].

Risks of Legacy Modernization:

In his book, Brook has identified the four measures of building software: *"(i) complexity, (ii) conformity, (iii) changeability and (iv) invisibility"* [51]. Legacy systems fail to achieve the two (complexity and changeability) among these four measures. It is due the fact that legacy systems suffer from the problems of rigid structural design, absence of proper documentation and lack of experts.

According to Geetha, the modernization of systems suffers from two kinds of challenges, which she observed during integration of systems. She has divided the challenges into two parts, one being technical and the other covering the human issues (non-technical). She said, "*The technical part covering, Usability, Software Development Service and Support, Security, Data Migration, Code Maintenance and Management, Strategy for Developing Migration Process Success. From non-technical side, the challenges are more from human factor such as Fear of the new software,*

Knowledge is power, cost of training personnel for the new tools, reduced productivity of the personnel". [52]

The legacy systems are ubiquitous and interdependent. Thus, the change in one system demands the change in other. This is one of the major challenges of modernization of legacy system and can causes difficulty in the working of organizations [37, 27]. Due to the complex structures of legacy system, it becomes difficult to modernize them. As legacy system suffers from compatibility issue with other systems. They are less receptive to other software and make the integration with other systems insignificant [37].

The legacy systems are written with old programming languages, which are difficult to interpret nowadays. These languages are difficult to reestablish, as their codes are only known to a few experts. In addition, to preserve these languages is also a challenge as they are not updated or documented properly. Hence, it becomes immensely problematic for companies to extract any valuable information (data and codes) from the legacy system for modernization [53].

In summation, all the aforementioned risks attached with legacy system modernization make the organizations reluctant to adopt modernization policy. Mostly experts do not want the legacy system to be replaced or modernized. The survey conducted by CIO Insight Magazine in 2002 has listed the following reasons for resistance of modernization:

1. Documentation of legacy system.
2. The cost incurred in professional training and replacement of system.
3. Difficulty to switch in terms of pausing the company's operations.
4. The risk of bearing the flaws of replaced or new system.

Lastly, not only the challenges discussed above makes the modernization difficult but also the culture of the organization has a greater influence on this policy. The study conducted by Xia & Lee in 2004 has shown that the major hindrance in adoption of new software is not the technical complexity rather it is about the perception and culture of that organization [54]. Hence, the organizational aspect has far more weightage in making the development of software successful than the technical aspect [55, 56].

Methodology of Legacy System Modernization

Though there exist several legacy modernization techniques, yet their implementation varies organization to organization. Various factors have to be taken into consideration while applying these modernization approaches. For instance, a company has to be aware of legacy system's hardcore architecture, financial constraints, return on investment, legacy system compatibility with the new system, etc. Hence, modernization is not only a technical phenomenon but a business phenomenon as well [55]. The modernization is not only about finances but it also revolves around the procedure through which the modernization process is done. According to Seacord, Plakosh & Lewis the modernization process comprises of *"market forces, business strategies and prudence approach that*

outline a total project benefit based on cost, benefit, risk and flexibility" [57]. The following theory will explain some of the widely used approaches of legacy system modernization in detail.

I. Integration of Legacy Systems

In the global era of constantly emerging competitive market, the availability of non-erroneous data is essential. Mostly organizations use legacy system for this purpose. However, they face the issue of not being at par with the latest technology. Hence, an application system like them is required to be restructured in order to make them compatible with the advanced technology. To meet this need, the legacy systems are modernized in such a way that they are allowed to integrate with other systems.

One of the key factors of integrating legacy system with any distributed enterprise system is their compatibility. This issue has been addressed in Enterprise information systems: 8th International Conference in detail [25]. The ability of legacy system to adjust with the newest technology or interoperate with other systems is the main analysis of The Integration and Interoperability Issues of Legacy and Distributed Systems [26].

In addition to that, integration of diverse systems also requires a proper programming language code [24]. Therefore, it should be ensured that semantic model for integration is used. As the manufacturing industry is changing continuously due to continuous modification in the information flow, dynamic integration of legacy system is required to make the legacy systems stable.

Different approaches have been presented for the integration of legacy system. Conversely, these approaches are also changing quickly to meet the requirements of the changing technology. Few approaches are definitely on the verge of decline. Firstly, because of huge investment of maintenance of legacy system and secondly, due to less chances of evolution of existing traditional system even after integration. [23]

The increase in demand of readily available data in the manufacturing industry has led to integration of multiple data sources. The data collected from these diverse sources is ought to be organized, scrutinized and adjusted. For this purpose, a platform, which can proficiently manage the data coming from diverse sources, is needed. An IEEE journal titled as *'Distributed control application platform-a control platform for advanced manufacturing systems'* has discussed in detail about the potential platform [38].

Likewise, a lot of work has been done on the integration approaches of the legacy systems. Some examples of integration include: agent based wrapper procedure, Business Process Reengineering, migration of data and Role Based Access Control [39, 40, 41, 42].

II. Data Migration

Modernization of legacy system includes data migration as well. The data plays the most important role in the legacy system, hence, it needs to be formatted and structured in an efficient way so that it would be easy to cope up with the new/other systems [34]. The restructuring of data comprises

right from designing the tables merging and mixing them to normalizing the data in a proper way by adjusting them.

III. Reengineering of Legacy System

Whatsoever modernization technique is adopted, it is essential to be fully aware with the architecture and coding of legacy system as well as the new system [55]. One of the most crucial features in understanding the legacy system modernization is the business value of the legacy system. It is a pre requirement of modernization so that the new target system can be built on that information. Aversano &Tortorella have elaborated in detail the need of knowing the business value of legacy system prior to modernization in their research paper [50].

Reverse reengineering is the process through which maximum knowledge can be extracted from the existing system [46]. Thus, the understanding of the legacy system gained through reverse reengineering is essential for modernization so that the needs for restructuring can be identified and the requirements can be met accordingly [58]. Usually this process is done by examining the smaller parts of the system.

The main idea of the reverse reengineering is either to build a new system that will be able to meet the requirement or to maintain the existing system by formatting it in the most suitable way [28]. This methodology gives the critical structural details of the legacy system [59]. Therefore, it becomes quite easy to know what kind of business value resides in that particular legacy system whose reverse reengineering is done [35].

It has been discussed in detail that the source of erroneous data collection is the presence of legacy system. The outdated system with its rigid structure and resistance to change according to new technologies and market needs, made the organizations to work on legacy system replacement or modification.

Now, this thesis will give an elaborated view of implementation of industrial systems in the organizations. In short, the subsequent section will explore the need of automation or technology, and the importance of legacy system in the industries/organizations. Further, it will give an overview on how huge amount of data led the companies to switch to automation and thus opt for application of legacy system.

2.3 Industrial Systems

Industrial control system (ICS) is a general term that encompasses several types of control systems, including supervisory control and data acquisition (SCADA) systems, distributed control systems (DCS), and other control system configurations such as Programmable Logic Controllers (PLC) often found in the industrial sectors and critical infrastructures. [17]. The current challenges of businesses demand high quality, increased productivity, low operating cost and better safety, which have drawn the attention of industrial manufacturers towards a more integrated solution. The role of technology in handling these challenges and coming up with a holistic solution has been excep-

tional. With the evolution of technology, the industrial systems have transformed from mechanization to automation. Hence, industrial automation systems have become the optimal solution since the last few decades giving more accuracy, productivity and efficiency. They give exceptional performance by replacing the labor-intensive manual machinery/systems and processes with computerized/automated ones.

These automation systems generate a huge amount of data for reporting, storing, managing and other production purposes. This data is collected through various manufacturing processes and automation equipment. Thus, the accuracy and reliability of data along with its easy access is very essential. As, automation equipment requires proficient acquisition of data, so, it has to be trustworthy. In case of faulty or inaccurate data, the complete automated industrial production is forced to stop working.

2.3.1 Automation in Industrial Systems

Automation systems and applications are pervasive in the emerging era of latest technology. The spread of automation technology is so widespread that life is not imaginable without them now. Implanted automation has so many functions with minor detailing, for instance, a car has so many functions, which keeps the car running and makes it a safe drive too. Automation systems are used to in the production as well as in the process industry. They provide the system architecture that is used in all the stages of production life cycle. [83]

Automation systems may consist of single programmable logic having sensors and actuators or it can be as widespread controlling many different systems and applications. They may be used to control few parts of the production lifecycle or some parts. Automations systems processes are found in all the three systems of Enterprise Resource Planning (ERP), Manufacturing Execution System (MES) and Distribute Control System (DCS).

2.3.2 Objectives of Automation

There are several objectives of automation. The following mentioned reasons of automation allow the user advantage over their competitors. [84]

Maximizing Efficiency

The automation system allows the user to have better productivity. It means that work can be accomplished by having few labor and more machines. The automation machinery also helps in providing more flexible production schedule, which operates day and night without any delay.

Improved Product Quality

Automation machinery allows better quality of products. It enables to manufacture homogenous products with fewer defects. It means that there is a consistency in product quality as compared to

manually manufactured product. In this way, the chances of having errors are minimized by automation machinery manufacturing. In addition, it eases out the later stages of production process.

Efficient Stock Management

Automation machinery has revolutionized the production process. The time taken to manufacture a product is reduced and the quality of the product has improved. Along with that, it has allowed the manufacturers to have smaller stock. As the ordered are placed online, the inventory record is maintained on systems, the delivery becomes convenient. The forecast of inventory in stock and on demand is properly administered and hassle of keeping large inventory stocks is reduced. In addition to that, the problem of handling outdated or expired inventory is also catered, as the order is placed when the demand is raised.

Information Management

Due to availability of data of manufacturing and inventory system, the customer has been kept in loop about the different stages involved in production, displacement and delivery of the product. The availability of information to both producer and customer makes the SCM process proficient and smooth.

Secure Production

The automation machinery enables the production process to be safer than manual production. For instance, automation system can send out alarming signal if there is a discrepancy in the safety of the plant or worker. The chances of hazardous incidents are reduced and the safety of workers is ensured by deploying automation system. Hence, the productivity of employee is enhanced and the distractions in the production process are reduced.

Keep Track of Defective Products

The availability of data in the automation system allows the manufacturers to keep track if any defective product is returned. It facilitates the process of ensuring quality and adjusting changes in the product according to the demands of the user. It is possible because the production number and batch id of the product is recorded in the automation system.

Prediction of Defects

Installment of automation machinery has given room to predict the malfunctions and defect in the system prior to facing any consequences. The automated product manufacturing has a monitoring system too, which diagnose about the existing failures and warns about the predicting malfunctions or errors in the system or products. In this way, the production system runs competently without any potential breakage or gaps.

In sum, there are compelling reasons of revolutionizing the systems from mechanical to automation system. One of the most important one is the information management because first it allows managing the enormous of data efficiently. Second, it keeps the user updated. Keeping that in view, the

following section will give a synopsis of the journey from data collection to information management. Though detail of it has already been discussed in 2.1.1, the subsequent theory explores the role of different automation applications in information flow management. More specifically the flow of data in different levels of ISA 95 Standard will be discussed.

2.4 Information Management Systems

Nowadays, companies are looking for competitive advantage to enhance their global and local market share. For this reason, Information Communication Technology (ICT) is used for efficient information sharing and communication to improve the client and manufacturer relationship. Hence, information systems encompassing the whole operations of a company are required to increase the flow of information between different production levels and stakeholders. The following three information management systems (Supply Chain Management, Enterprise Resource Planning and Cloud Computing) are the product of latest technology to achieve the desired goal.

2.4.1 Supply Chain Management

Supply Chain Management (SCM) has been an important concern for businesses in order to have an efficient platform for production and delivery of products and services. In addition, the integration of suppliers and users has been a major worry in the supply chain. SCM is defined as *"the integration of key business processes from end user through original suppliers that provide products, services, and information that add value for customer and other stakeholders."* [98]

SCM is an approach, which streamlines the production and distribution process in order to increase the company's competitive advantage and performance. It is a process, which integrates the flow of information from supplier to distributor. The cooperation between the user and manufacturer are essential for the best optimization of requirement of product. This cooperation can only be ensured if there is a platform for real time information or data sharing platform between the supplier and customer.

The quality and standards of information shared between the user and supplier carries much importance. The information should be reliable, timely, accurate and consistent to increase the performance of supply chain. This information sharing helps to mitigate the supply and demand uncertainty in the supply chain. Since long time, the companies have been looking for technologies to enhance this information-sharing platform in SCM [99].

2.4.2 Enterprise Resource Planning (ERP) Systems

One of the reasons of highlighting the importance of data quality and information flow is the need of getting an integrated view of data and information for decision-making. From data integration, one can infer that it is the process of combining two or more data sources to get the information needed and to improve the data quality [91].

As information systems are the backbone of providing concrete information for proficient decision making and operations management, the organizations are investing in replacing the legacy systems with Enterprise Resource Planning (ERP) systems [92]. Thus, keeping the importance of information systems in view, ERP systems are a reliable solution to get a broader and better support to business activities.

The main function of Enterprise Resource Planning system is to enhance and boost up the flow of information in the organization [93]. From inventory management to human resource management and from marketing to finance, the role of ERP software is to allow smooth information flow and information sharing. All the business units of the organization communicate easily with the help of ERP software implementation in the organization.

In addition to that, ERP allows to gather standardize information from the data. This in turns harmonizes the data flow within the different units of the organization [94]. The essential features of ERP are to integrate the information in such a way that it speeds up the processes and makes them cost effective. The data sharing is made real time for comprehensive view of all the available resources within the organization [95].

ERP in SCM

The integration of ERP in SCM has been the widely choice solution opted by various companies to consolidate their supply chains [100]. It allows a smooth planning and operational platform for SCM. The following are some of the main benefits of ERP in SCM:

I. Enhancement of Supply Chain

With help of ERP implementation all the activities of supply chain; from raw materials to production and from production to delivery are enhanced. All the operation of supply chain can be visualized from a real time platform of ERP.

II. Increased Optimization

One of the main challenges of supply chain is the timely delivery of products. ERP allows the delays in distribution and delivery of products through its optimization functionality. It also improves the demand forecasting in SCM.

III. Information Sharing

ERP bridges the gap between the supplier and customer by allowing streamlined information sharing between them. The whole network of supply chain is benefited from the inventory management, demand optimization, status of production, and transport arrangement through real time information sharing by ERP. This kind of collaboration between the different actors of supply chain allows smooth flow SCM operations [101].

IV. Cost Effectiveness

ERP enables cost-cutting opportunities by providing demand forecasting and inventory management. The need for storing extra products or raw materials is reduced by ERP. In addition, the customer needs can be satisfied without any shortages through ERP [102].

2.4.3 Cloud Computing

Due to the increase in demand of supply chain management, there is a need for more collaborative platforms, which can ensure timely and efficient delivery of products and services. The uncertainties in the market and economic crises have forced the organizations to look for more efficient techniques to keep the businesses running in a cost efficient and timely manner.

Cloud computing consist of network of servers to provide the storage, monitoring and collection of data over the internet. In supply chain management, cloud computing is responsible for giving the overview of the product throughout the different stages of product lifecycle.

Cloud Collaborative Manufacturing Network (C2NET)

C2NET is a project which works on the cloud based computing methodology [16]. It provides a platform to support the supply chain management. It gives a real time data collection framework for real time decision making through collaboration and optimization tools. It allows improved data security and optimization. The scope of this project is to support all the stages of supply chain manufacturing; from manufacturing to delivery.

This thesis has been a byproduct of C2NET project. The project involves a framework for data collection from different legacy systems consisting of various protocols and databases. It facilitates in providing a seamless platform for the integration of system.

Data Collection Framework (DCF)

The DCF collects real and raw data from various diverse products. It gathers and then classifies data followed by detection of any discrepancies. The following are the features of C2NET data collection framework.

I. Interoperability

The compatibility or interoperability of the systems is required for integration of systems. Hence, C2NET ensures flawless compatibility between systems without the support of any external device or tool.

II. Adaptability

C2NET Platform allows the legacy system to adapt themselves for any changes without changing the existing data.

III. Security

The problem of data security against malicious data and other cyber-attacks has been solved by C2NET platform by giving reasonable solutions. Cryptography, secure connection and digital signatures are among those few solutions.

Legacy System Hub (LSH)

In order to collect the enormous amount of data, DCF is joined with Legacy System Hub, which gathers the data from Legacy System sources. The Legacy System Hub contains the following resources for data gathering:

I. Enterprise Service Bus (ESB)

ESB is an intermediate resource between C2NET Platform and Legacy Systems.

II. Simple Data Adapter (SDA)

The purpose of data adapter is to combine the data gathered from different Legacy System into a hidden format so that the information can be processed when required. A hidden layer conceals most of the technical aspects of data collection in C2NET.

2.5 Information Flow

In the past, organizations used to deal with the conventional methods of information flow handling which were time consuming as well as ineffective like manual data entry and the use of paper. With the advent of new technology more specifically after the Internet revolution the organizations like eBay and Amazon came up with the new system of information flow handling rather than revising or improving the older way. These post revolution organizations introduced the customer satisfactory method of information flow. For instance, eBay receives the order from customer through internet, it then process the information further for the prompt delivery of the order. In such a way input is fed (order is placed) electronically then processed, and the loop is closed after the acknowledgement of the shipment electronically. This shows an assimilated integration of internal processes of company when an external order is placed.

The data, which is generated by automation system, is collected through various functions associated with production. This data is utilized in numerous ways to improve the process of production and to maintain the quality of automation equipment. As it is discussed above that the quality of data is an integral part of data management, otherwise, drastic consequences can be encountered. Hence, it should be made sure that reliable data is accessible to several types of automation applications.

2.5.1 ISA 95 Standard

The synchronization of information systems has not been observed on the different information flow levels. The first developed information systems were established at the enterprise level in 1970s.

They were used for accounting and stock management of the companies. Later, the gradual development of enterprise level information systems allowed their use in manufacturing industry and production management. Through this, these systems were made dependent on the latest production information/data. The emerging Enterprise Resource Planning (ERP) systems are used to manage almost all kind of resources of a company, from marketing to human resource and from sales to financial resources. Nonetheless, ERP systems are more of financial monitoring software rather than looking after the production and manufacturing processes. This biter fact has generated a huge difference between ERP and automation process control systems. Data management problems have escalated because of this gap between the two systems. All the measures of data quality like timeliness, accuracy, consistency, etc. are a result of this gap. [72]

With the advancement in the global technology, the digital computer based systems have the possibility to spread into the process industry information as well as into the automation applications. However, these functions were not available when the automation applications started in the industry. Due to this, it is difficult to integrate and interface for the automation industry, the divergent systems used in the business processes.

The ISA SP95 Committee took the responsibility of integration of these two fields of automation applications. The objective was to come up with a standard through which different modules and hierarchy levels of information system can be defined. Moreover, it reduced the cost and errors during implementation of interface. In addition, it provided safe, efficient and reliable interface. Lastly, it allowed maintaining the data integrity. [74]

2.5.2 Hierarchy Levels of ISA 95 Model

The ISA 95 Standard consists of functional hierarchy level, which represents the flow of information in the manufacturing industry. Each level has been elaborated in the Figure below and shows the functions related to it.

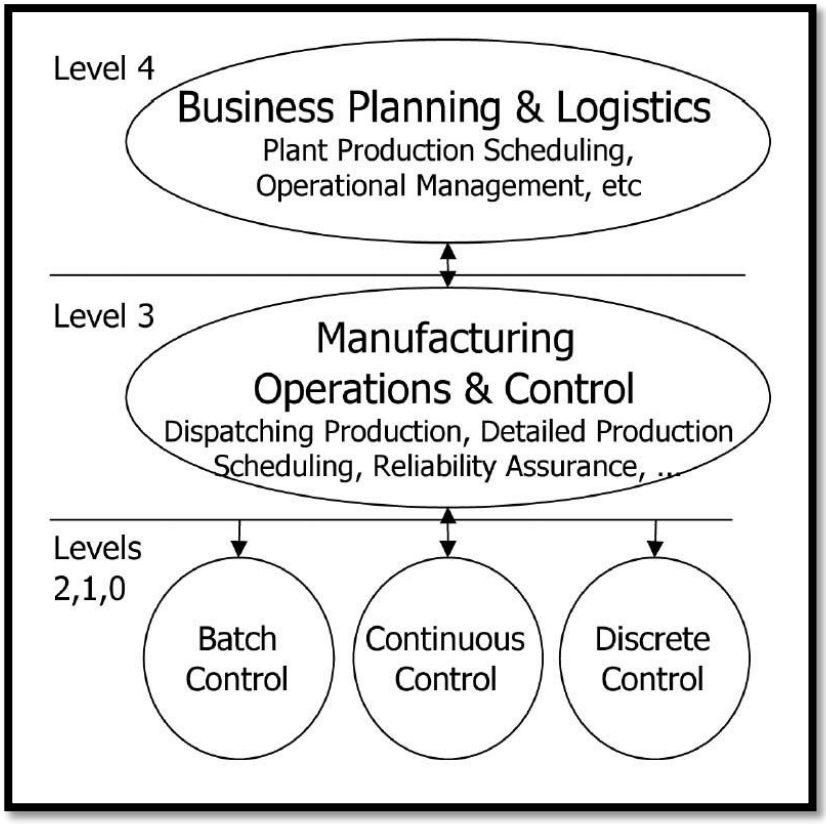


Figure 6. *Showing the simplest version of functional hierarchy model by Bianca [73]*

Automation applications can be built on the ISA 95 hierarchy model. The ones located in the lower levels (levels 1 and 2) of the model are responsible for maintaining the production, for promoting the visualized data from factory level (lower level) and for business functions related to data reporting. Whereas, the others, which are located in the upper levels (levels 3 and 4), control the automation process and manufacture the process data. The data management of any automation system faces several problems allied with security services, configuration services, the backup of data and validations, audit trajectory and recording services [14, 15].

On this application structure, the flow of information can be seen from lower level to the upper level of the ISA 95 automation pyramid. The upper levels use various technologies and techniques to store the information provided by the lower levels, which include conventional databases, REST services and other data storing and managing tools like Excel sheets.

Concisely, the information runs through various these four levels of ISA 95 model and hence gives the user the knowledge, which is required. Moving on, the dilemma of transferring the information between two alienated automation systems is done by function blocks. They supplement the information flow management in such a way that they have the agency to acquire data from one system, convert it into information and transfer it into another desirable system. The next section of the thesis gives an all-encapsulating view of few function blocks and their role in integration of legacy system hence, information flow.

2.6 Function Blocks

This section will discuss the concept of service-oriented architecture and what role it plays in the implementation of the IEC 61499 standard. The concept of function blocks is then discussed in the scope of this standard. In the end, this section will give a brief introduction of the PlantCockpit project and talk over the already implemented data adapters.

2.6.1 Service Oriented Architectures (SOA)

Due to the interdependency of the internet devices, it is required to have a flexible, independent, communicable and affordable device, which can connect various dependent devices without plug and play. With the increase in the time efficient and time oriented manufacturing industry, a proactive interface is needed which can help in configuration, maintenance, error inspection and examination of the processes. [75]

The Service Oriented Architectures (SOA) is the manufacturing systems design, which is interoperable and autonomous. These two features of SOA systems make them best architectural standard in theory. They are autonomous in a sense that they work independently. In addition to that, they can process smoothly without being connected to any advanced system. [75]

On the contrary, SOA makes the system interoperable. Interoperability is supported by plainly abstracting the interface that an administration opens to its surroundings, from the execution of that administration. Nonetheless, the reunion of these two opposing features is one of the challenges, which SOA faces. [75]

Initially, SOA used to be the company's rule due to its involvement in business processes. Later on, it was developed and used for interoperability of different domains through a single platform. Hence, SOA features include scalability and evolvement as well [76]. The deployment of SOA in automation system is not a new concept. Epple has presented this concept in the following figure also, which explains the working of SOA in automation pyramid [77].

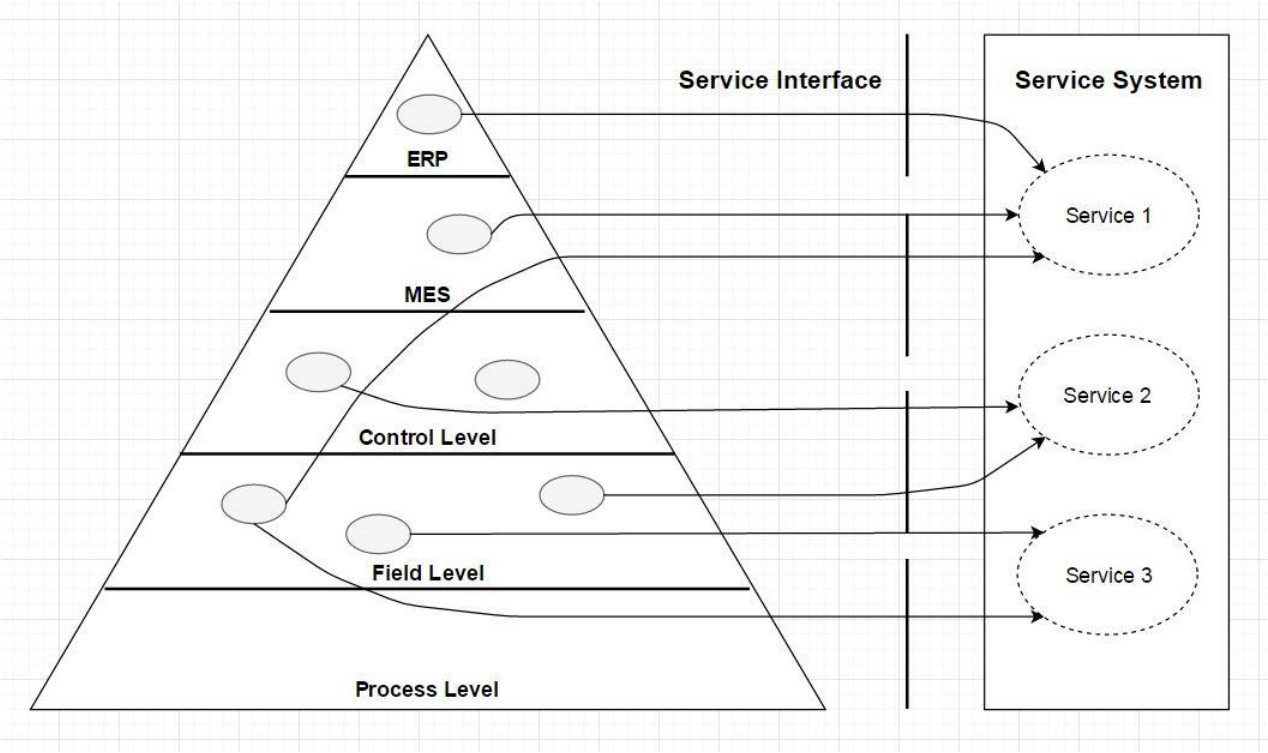


Figure 7. *Describing the relation between SOA and automation pyramid [77]*

2.6.2 IEC 61499 Standard

The IEC 61499 is defined as the new standard or language for distributed control systems [79, 80]. It is a revised version of the IEC 61131 standard, which handles the Programmable Logic Controllers (PLCs). The IEC 61499 is a form of Functional Block. Generally, Function Blocks are referred to a concept of diverse independent automation units and their interaction with each other. The defining feature of IEC 61499 is that it allows the user to have more specific information about the computational units as compared to its ancestor IEC 61131.

The Function Blocks enables the user to have more vigorous and easily accessible data units [79, 81]. Every Function block has two major parts, which are described more fully in the figure given below:

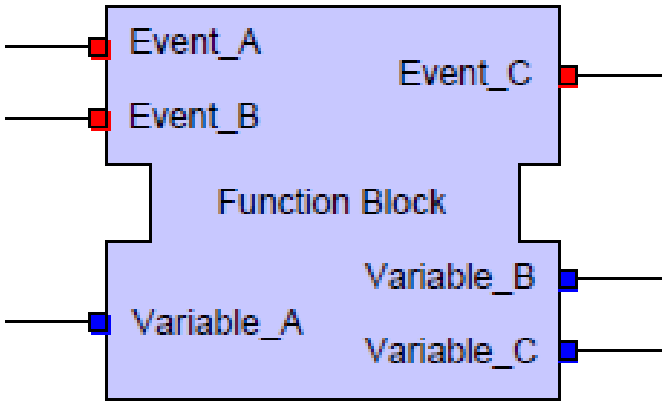


Figure 8. *Showing a standard Function Block [82]*

The one part of Function block is the controlling part while other is about data. The input events, which are introduced into the Function Block, are mentioned on the left side of the figure. The task of the controlling part is to monitor them. It depends on the condition of Function Block to send them outside or retain them. The computational results allow the Function Blocks to enter the input events from left and to send data, which can be reused through output units on the right to other Function Blocks. The interconnected Function Blocks have been illustrated in the following figure. [82]

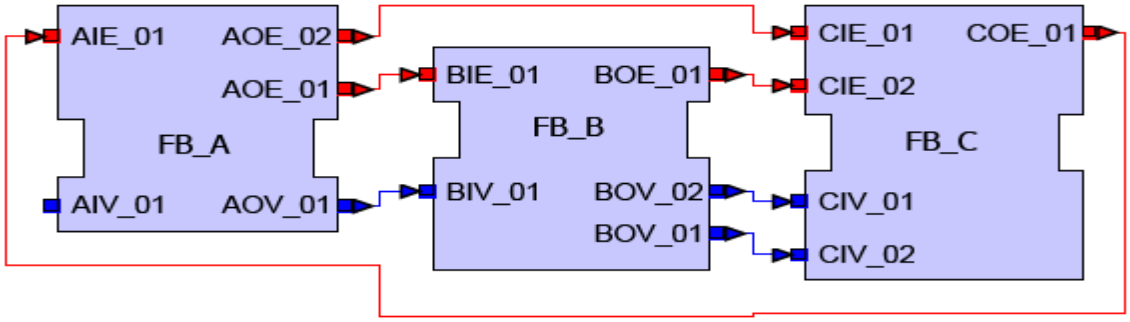


Figure 9. *Illustrating the network of interconnected Function Blocks [82]*

PlantCockpit (PCP)

PlantCockpit (2013) project is implemented to mitigate the challenges in integrating two alienated systems by acting as a compatible bridge between them [16]. PlantCockpit use the function block approach discussed in the IEC 61499 standard and the service oriented architecture (SOA).

Previous Implementations

I. SQL Data Function Block

SQL Adapter is made to fetch data from relational databases with the use of 'Structured Query Language'. This Adapter will take input configurations from the user. These configurations are in

JSON format and contain three main parts: headers, sources and output schema. Headers have different ids to identify the created instance of adapter. While, sources have the data related type, location and authentication of the relational data base. Further, the output schema contains the structure of the JSON output required by the user having all the information for creating a query to fetch particular data to replace in the output schema. This adapter is created as such to accommodate different type of databases. The current version can be used to operate with MySQL, MariaDB and PostgreSQL.

II. Excel Sheets Data Function Block

Excel Adapter is made to fetch data from different Excel files likewise. These Excel files are hosted on an ftp server which can be used to fetch these files remotely from anywhere on the internet with proper authentication. Configurations in case of Excel adapter are principally same, the main difference is in the sources, where now there are Excel file name, FTP server and authentication details and in output schema in place of query details there are cell references to the excel file. This adapter also returns data in JSON format as specified by the user in the output schema.

2.7 Review of Theoretical Background

As major part of this thesis has addressed the previous theories and information available on the relevant issues which will be addressed in the thesis, this section will summarize the major highlights of Chapter 2.

Starting from data management, it has been observed that data, being the vital asset of the organization encounters problems with quality and reliability. This in turn distorts the information which the user requires from the acquisition of data. The later chapters of this thesis will bring attention to this problem of the data management.

Moving on, the issue of legacy systems data acquisition and the concerns related to modernization of legacy system posits about the major gap in the industrial systems. Similarly, there have been information management systems but the lack of one single platform to acquire data from multiple data sources of legacy systems is missing. Hence, the next sections will attend to all these gaps.

3. RESEARCH METHODOLOGY AND MATERIALS

This Chapter gives the details about the research objective of the thesis. The research questions which have been formulated, and the approach, which has been adopted to carry out the research in this thesis. It gives an exhaustive study of the already available research methodologies and approaches and gives an overview of the research model designed for this thesis.

3.1 Derived Research Questions

The research of this thesis revolves around the data collection framework in industrial systems. As mentioned earlier, the problem lies in faulty quality of data, which is acquired through legacy systems. Thus, the thesis will delve deeper into the following research questions:

1. How data acquisition adapters can mitigate the challenges of coping with legacy systems?
2. How information flow can be enhanced to retrieve the readily available data required by user?

The following figure depicts the research model of this thesis:

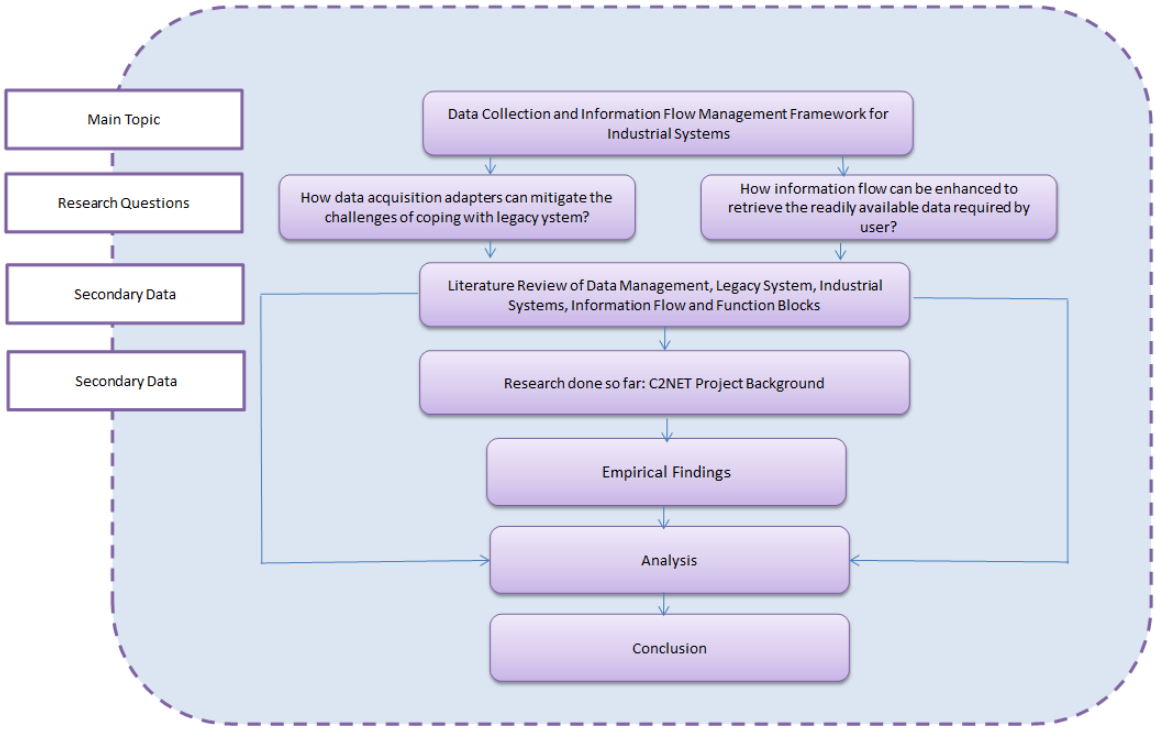


Figure 10. *Research Model of the Thesis*

3.2 Research Phases

The work presented here has been gone through different phases of analysis, brainstorming, proof-reading and collection of relevant content. The research material gathered was unorganized in the beginning and was managed according to the structure after going through different phases. The first four phases were the most critical in the structuring the thesis in order. The following measures taken to address the research question will provide an overview of the process through this thesis has been gone rather than the sequential steps taken to formulate it:

3.2.1 Initiation and Planning

It is quite difficult to anticipate how the research will go about from the beginning until end. Though the research questions and the guidance of authors help to formulate the basic order, yet a good amount of investigation and examination of the research was required to delve deeper into the heart of the thesis. The planning also helped in deciding the approach to follow in the thesis.

3.2.2 Theoretical Foundation

A major chunk of the thesis comprises of theoretical research. This was done in order to lay the foundation of thesis. It comprised of secondary research; articles, journals, reports, etc. Lot of literature has been read during this phase to get the gist of the research problem, previous work done and relevant data available. This phase actually helped in finding out the gaps, which need to be addressed. All the theoretical background has been separately discussed in Chapter 2 of the thesis.

3.2.3 Problem Identification

After going through a vast amount of literature and addressing the basic concerns of data collection framework and information flow in industrial systems. The areas, which need to cater, were identified. Hence, the research problem presented in section 3.1 was formulated. These research problems/questions form the basis of the topic of discussion of the thesis.

3.2.4 Empirical Investigation

The next phase of the thesis was to look for working hypothesis. This was done by conducting test and figuring out the results. The description of which has been addressed in chapter 4 of the thesis.

3.2.5 Development of New Research Endeavors

The whole process from collection of data to reviewing the literature to going through empirical investigation generated new dimensions of research. Few areas, which were identified, has been addressed in the thesis while the rest have been still open for discussion. This thesis became a pillar to give theoretical background to these future research opportunities.

3.2.6 Documentation of Work

The final output or phase in formulating this thesis was the documentation of all the information in the form of dissertation. All the research, which was conducted, was drafted in the form of thesis in this phase and was gone through further analysis for outcome.

All these six phases provide an insight of all the procedures through which the thesis has gone through. The next sections of this chapter will reveal the approaches and methodologies to conduct research.

3.3 Scientific Approach

This part of the thesis will focus on the scientific methodology used to verify both the discussion and the test done to achieve the desired outcome. The first step of scientific methodology is to identify the problem, which according to this thesis was lack of efficient data collection framework that troubles the information flow. It is then followed by building a hypothesis, testing it and in the end analyzing the results.
Both science and scientific knowledge are the concepts, which have been debated a lot. However, scientific method is the approach on which there is unanimous agreement. This approach has always been used in history whenever there is some research done on a new paradigm. This thesis has encapsulated the two main objectives of the scientific approach; testing and outcome.

3.3.1 Inductive, Abductive and Deductive Reasoning

There are three methods of reasoning, whatever the approach of scientific method is adopted. These are inductive, abductive and deductive reasoning. Both inductive and deductive are based on some prior information. However, the fact which differentiates both is that inductive relies on making assumptions about the past occurrences or information. Whereas, deductive draws conclusions and inferences on the already available theories and documentation. The distinction may not be as sharp, but becomes quite evident while we consider these reasoning approaches from re-search point of view. For instance, while conducting research, deductive reasoning is based on the hypothesis formed by the prior theories and information, and all those theories and data are rendered significant if the hypothesis turns accurate. On the other hand, during inductive reasoning, the interpretations and analysis are the basis of forming new hypothesis and further theories. [71]

To define the abductive reasoning is a bit controversial process. It is an approach, which is formed while studying the already available theories. It is somehow a reasoning, through which an already verified hypothesis is taken into consideration as contrary to forming a different one. It is also considered a combination of both inductive and deductive reasoning. This means that in abductive reasoning, first the observations are verified through deductive reasoning and then theories are formulated in an inductive manner. There are plenty of ways to define this concept. Some consider the abductive reasoning as the only approach, which generates new information. [71]

In short, in the thesis mostly the approach used to test the hypothesis is an inductive one, based on observation. Nonetheless, abductive reasoning can also be found, while considering the new result that has been generated at the end. Still, nothing is adamant regarding the reasoning approaches used in this thesis and there can be found traces of each one of them according to their description presented above.

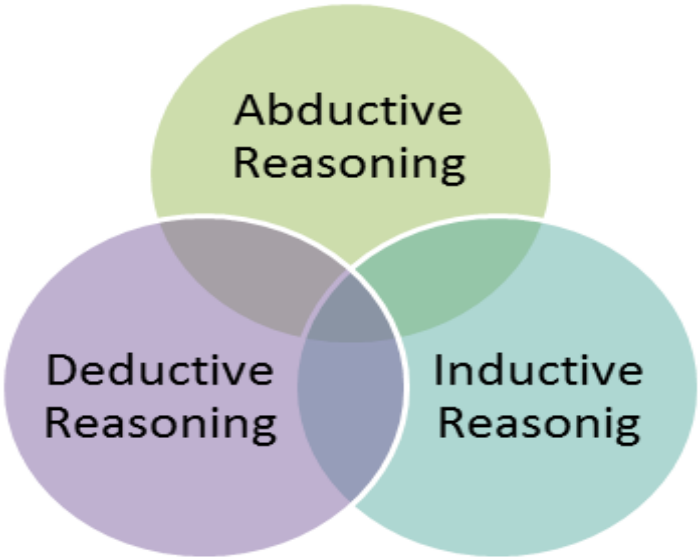


Figure 11. *Forms of reasoning*

3.4 General Approach

The approach of this thesis encapsulates the main components and tools used to achieve the solution of the defined problem. The interaction and communication between these components, their function and their role in providing an overall view of the framework has been elaborated.

The main objective of this thesis is to redesign and modify a user convenient data collection framework. Thus, the DCF of C2NET project has been the basis of achieving the set target. The data collected in C2NET platform is mainly from ERP systems deployed in companies. Hence, the ERP data acquisition and retrieval is the main goal of C2NET. The outstanding feature of this methodology is that a wide variety of dissimilar data is collected from heterogeneous sources and is harmonized in the process.

Since, RESTful web services, which are the mediums of access for the user to the resources, are one of the significant sources of data, thus, they have been discussed in finding the solution of the research questions mentioned in Chapter 3 of this thesis. Apart from that, SQL data sources, which are used for retrieving and updating the data from various relational databases, have been analyzed. Lastly, MS Excel files are used mainly in the companies, therefore, a major data is collected from them. The MS Excel data sources are one of the important data sources which have been projected in the below mentioned part.

4. EMPIRICAL RESEARCH

In this Chapter, the approach to achieve the set objective/hypothesis of the thesis has been discussed. The tools and techniques used in the development of the framework have been stated. This chapter concludes with the detailed description of the implementation of the approach adopted in designing the framework.

4.1 Tools and Techniques Used

Data Sources as Legacy Systems

As discussed in Section 2.2.1 legacy systems face certain issues regarding their incompatibility with other industrial systems. In this research the following data sources are categorized as legacy systems and will be used as use cases to prove the hypothesis:

1. SQL Databases (Data stored in tables with relation to each other)
2. REST Endpoints (Http endpoints returning data in JSON format)
3. EXCEL Sheets (Data stored in simple tables for keeping and presentation purpose)

***PlantCockpit* (PCP)**

PlantCockpit is a compatible integration open resource. It provides a platform to create Function Blocks like SQL, XML and REST Adapters. It is positioned above the ESB as part of the Legacy System Hub in the C2NET platform. PlantCockpit allows smooth integration of systems to maximize business control over supply chain and gives an unbreakable view of supply chain. The figure below represents a basic model of what PlantCockpit Adapters signify as part of a supply chain.

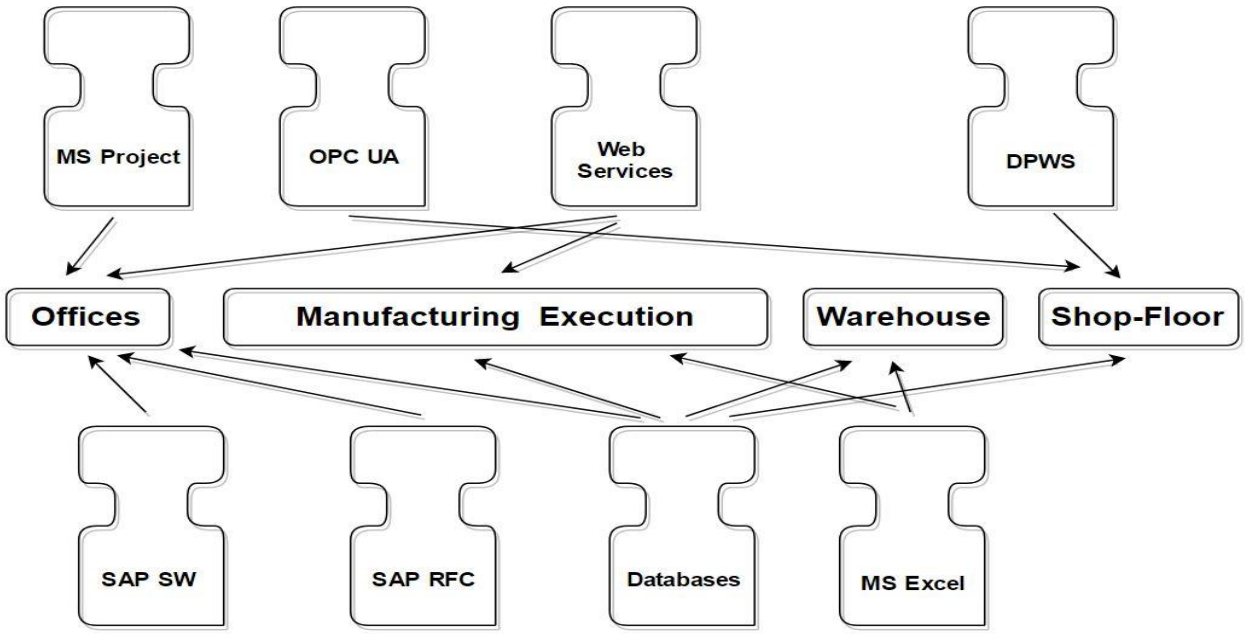


Figure 12. *PlantCockpit implementation of loosely coupled Adapters [90]*

PCP works on the IEC 61499 standard of Function Blocks. In this framework, its function blocks are used and their functionality is extended to create data system adapters. PCP provides loosely coupled components, which communicate with each other using interfaces. Flexible mapping of different sort and type of data can be easily done by communication through standardized XML message format channels. PCP also provides a user-friendly graphical user interface for the user to configure different function block according to his needs. That configuration process is designed in a way to provide the user best possible way to implement what he needs. It similarly provides highly extensible functionality, which can be enhanced or modified according to the needs of the user. [90]

Apache Service Mix

Apache Service Mix is an open source ESB founded on the principles of SOA. It is flexible and can be easily installed. It is a run-time collection of components of SOA, web services or legacy systems integration services. It offers reliable messaging, integration methodologies and routing. It is compatible with JAVA and its applications. The framework of ServiceMix is based on OSGi (Open Services Gateway initiative), which is able to remotely manage JAVA applications.

Apache Service Mix is supported by Java Business Integration (JBI) standard, which gives an overview of the component, defines the messages and their interactions. Both JBI and ServiceMix transfer the converted standardized messages to the ESB. The bus then transports the messages to the target system. Thus, the data is harmonized in this process [96].

Besides the JBI, ServiceMix allows integration and connectivity of various other components, which include, accessing XML files, supporting HTTP and exposing integration patterns for legacy systems [96]. One of the important features of ServiceMix is to allow reliable messaging through ActiveMQ Message Broker, which has been discussed in the next heading.

In this thesis we are using Apache ServiceMix to deploy our PCP Adapters with supporting bundles. The ServiceMix makes sure that all the dependencies needed to make a bundle work are present. It controls the status of the bundles and can be used to start stop and restart each working bundle.

Apache ActiveMQ Message Broker

Message brokers allow the integration of different architectures or systems. Their ability to accommodate various data types is appreciated by client or server. Especially, they allow XML data to be easily transferred between various heterogeneous systems. [97]

Apache ActiveMQ Message Broker is the most commonly used influential messaging server or broker for ESB implementation (ServiceMix). It is again an open source for integration services and is known for providing '*Enterprise Integration Features'*. It is based on Java Message Service (JMS) and is coded in JAVA. [97]

ActiveMQ Message Broker is a bridge between client and server and allows more than one communication link. It is significant because of the flexibility it provides along with the ability to sup-

port large protocols like REST. In addition, it is famous for supporting different kinds of Cross Language Clients and diverse Protocols. It provides outstanding client server, clustering, load sharing and smooth communication.

The ActiveMQ message broker communicates with all the components attached to it through an input and output topics. Each component receives an input message through its input topic and delivers its output to the output topic.

VMware Workstation

An Ubuntu virtual machine was used to deploy the ServiceMix for this project to ensure that this project can be run on generic platforms as a Docker image. The below figure shows the user interface of the VMware workstation used to host this virtual machine. The ServiceMix can be deployed on any Ubuntu machine using a Docker image.

Bitvise SFTP Client

To navigate through the virtual machine, manage and excess the files Bitvise SFTP Client is used. As .jar files had to be readily updated in the deploy folder of ServiceMix.

Software used for development and data management

For mocking and hosting the data for the use cases, the following software are used:

1. Eclipse Jee Neon for the development of the data adapters.
2. JetBrains WebStorm version 2.2 is used to host the mock data of REST sources for the REST adapter and managing code for the Resource Manager mock implementation.
3. MySQL Server is used to host the database for the SQL adapter.
4. MySQL Workbench version 6.3 CE is used to manage the MySQL Server database.
5. Microsoft Excel 2013 is used to create and manage the Excel files for the Excel adapter.
6. Postman client to check the HTTP endpoints created for the REST adapter.
7. Notepad++ for managing the JSON files for the configuration of the adapters and as an HTML editor.
8. Google Chrome as a default web browser to run the RM, ActiveMQ web console and any HTTP endpoints needed in the implementation of this project.

4.2 Implementation

4.2.1 Architectural Solution

The solution to the above discussed research problem is to provide the user readily available data in the form that he wants. The technical implementation of the solution can be visualized by the system

architecture diagram below. This diagram shows which components are involved in the solutions and what role do they provide towards achieving the desired outcome.

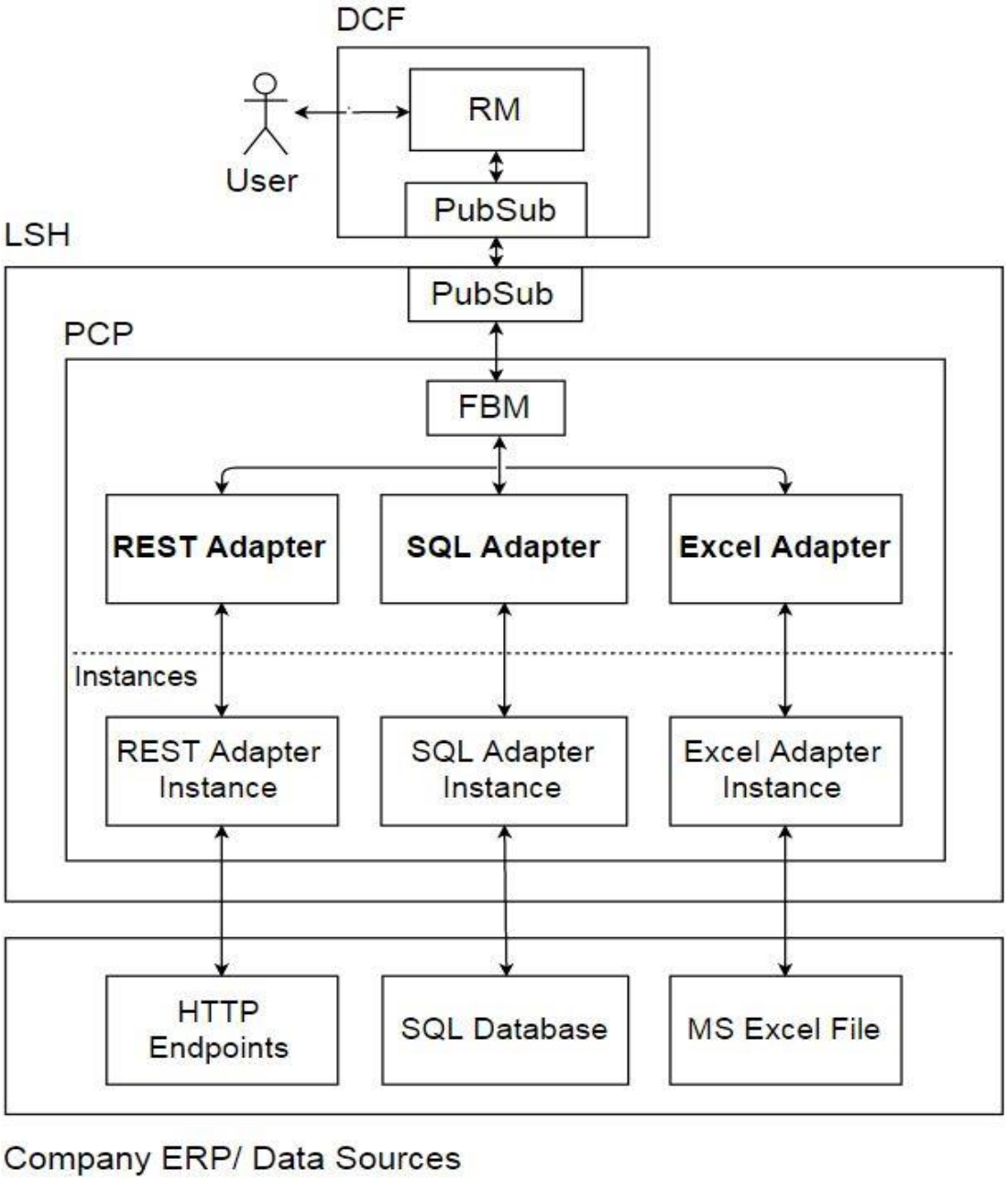


Figure 13. *System Architecture Diagram of the proposed solution*

According to the figure above, three main components are involved in the whole process: i) Data Collection Framework (DCF), ii) Legacy System Hub (LSH) and iii) Data Sources. The DCF includes the Resource Manager (RM) user interface, which is used to enter the fields for configuring the Function Block Instance (FBI). It also includes the server side of the PubSub module, which is responsible for communicating with the LSH. The LSH contain the client side of the PubSub module as well as the PCP implementations of data adapters. The Function Block Manager (FBM) helps the

Function Block (FB) to create its instance by parsing the initial XML configuration shown in the figure below. FBM gives the required JSON configuration object to the FB that creates its instance according to it. The instances then communicate with the company side ERP to get the required data from them. The Data Sources consist of three different types of data for current problem scenario, i) HTTP endpoint, ii) SQL Database and iii) MS Excel file. These data sources are continuously communicating with the LSH to give them the required data. A detailed overview of this information flow has been explained in Figure 16 below.

```
<?xml version="1.0" encoding="UTF-8"?>
<createFbInstances>
    <fbInstance>
        <fbId>test</fbId>
        <fbPid>pid_eu.plant.Testing_Block</fbPid>
        <businessConfiguration>{"configuration":""}</businessConfiguration>
        <inputMessageTypes>
            <inputMessageType>
                <messageTypeName>String</messageTypeName>
                <transformations>
                    <transformation>
                        <newMessageTypeName>STRING</newMessageTypeName>
                        <transformationScript>STRING</transformationScript>
                    </transformation>
                </transformations>
            </inputMessageType>
        </inputMessageTypes>
        <outputMessageTypes>
            <outputMessageType>
                <messageTypeName>String</messageTypeName>
                <transformations>
                    <transformation>
                        <newMessageTypeName>STRING</newMessageTypeName>
                        <transformationScript>STRING</transformationScript>
                    </transformation>
                </transformations>
            </outputMessageType>
        </outputMessageTypes>
    </fbInstance>
</createFbInstances>
```

Figure 14. *XML configuration sent to the FBM*

The structure of a created FBI is as such. The input configurations first go to the *Main Adapter* class, which initiates the *Engine* class. After that, the *Engine* invokes the *Source Manager* class object to manage the input source fields. At this point, the Engine class can use the *Data Fetching* class object to obtain data from the configured data source. In addition, it can it use the *View Generator* class object to generate stringified HTML for the RM. The architecture of the FBI can be visualized through the figure below.

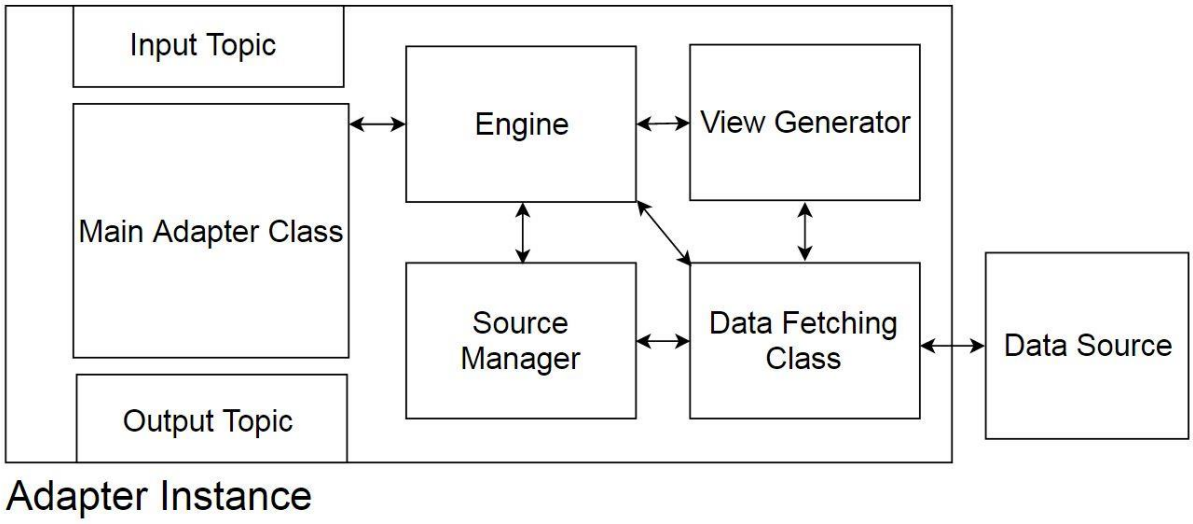


Figure 15. *Architectural illustration of an FBI*

4.2.2 Module Interaction

The discussion in the previous heading highlights the role of involved modules. This part is going to emphasize on how those modules interact with each other to produce the required output. The sequence diagram in the figure below (Figure 18) illustrates the communication done between each component of this project.

The user enters the configuration fields in the RM user interface. The RM converts these fields into half configuration by putting it in a JSON object. This JSON object is when received by the PubSub module is converted into a full XML configuration for the FBM. The FBM then creates the adapter instance and sends it the parsed JSON configuration. The FBI configures itself according to the user input fields, initiates the Engine and tells the PubSub that the configuration of the adapter is complete. At this point, the PubSub asks the FBI to fetch the column name fields of the configured data source. The FBI communicates with the data source and fetches the required information. The adapter then converts this information into an HTML string for the RM and sends it to PubSub.

The RM after getting the HTML string from the PubSub module; renders the HTML and shows it to the user in the web browser for him to select the fields for which the column data is required. The user by selecting the fields tells the resource manager to configure the LSH to fetch the data for those fields. PubSub on receiving this configuration from the RM parses it and sends it again to the FBI. The FBI again communicates with the data source to fetch the column data. After getting the information, it organizes the information in form of a JSON object and sends it to the RM as required by it. The FBI then starts checking the data source every five seconds for any change or new entries. If the data is updated in any way, it will again fetch it and send it to the RM. This loop keeps on repeating until the user destroys the FBI.

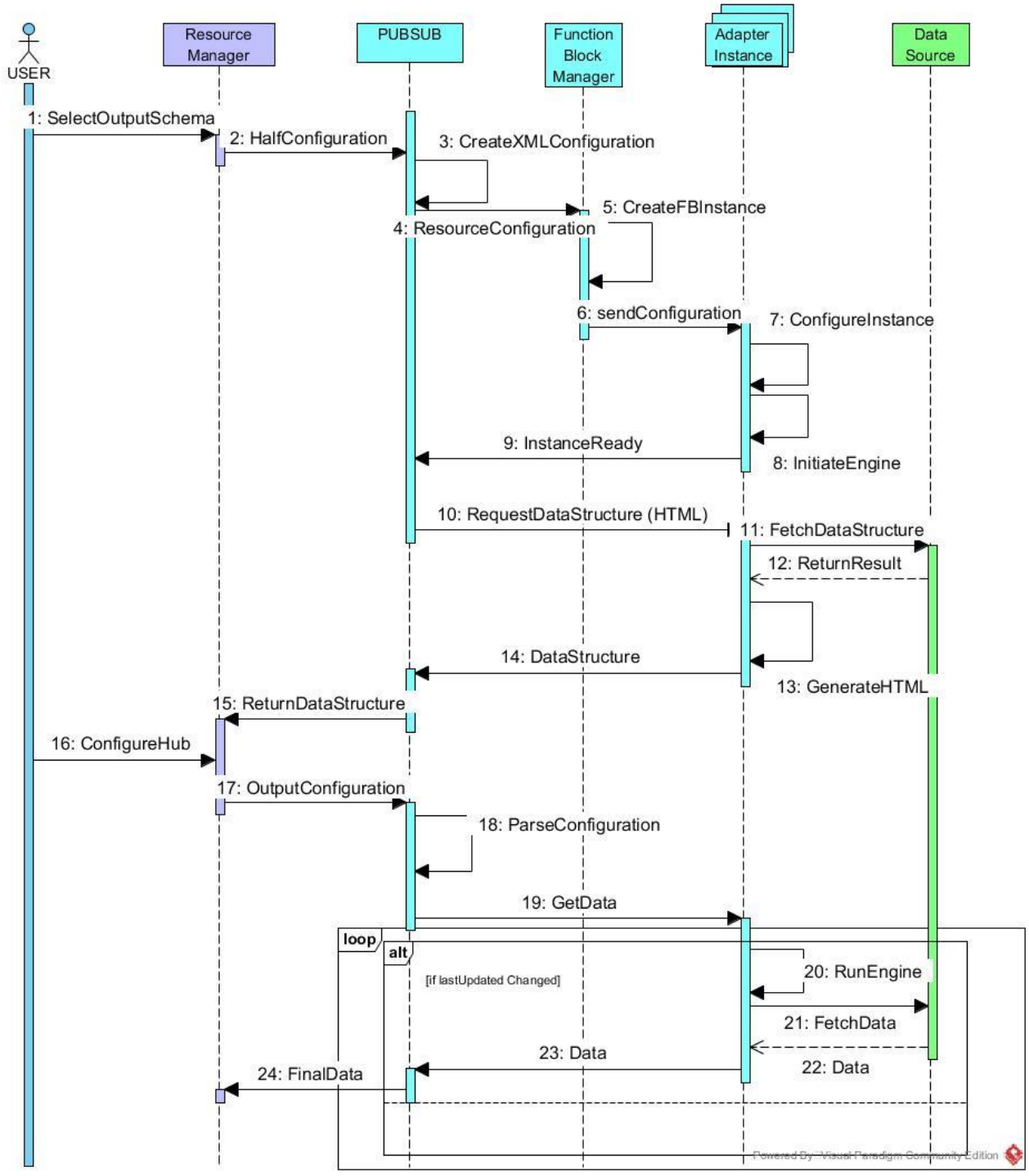


Figure 16. *Sequence Diagram showing the project module interaction*

4.3 Use Cases

This part of the thesis will discuss the use cases implemented on the base of implementation discussed above. Three cases of different types of data sources are presented. This section will emphasis in detail how each of them is constructed and what differences do they have from each other in their implementation.

4.3.1 REST Adapter

The REST adapter is created to acquire data from different REST endpoints hosted on different HTTP servers. These online servers return data in the form of JSON object back to the user (RM interface).The figure below shows a mock server implementation of one of these REST sources.

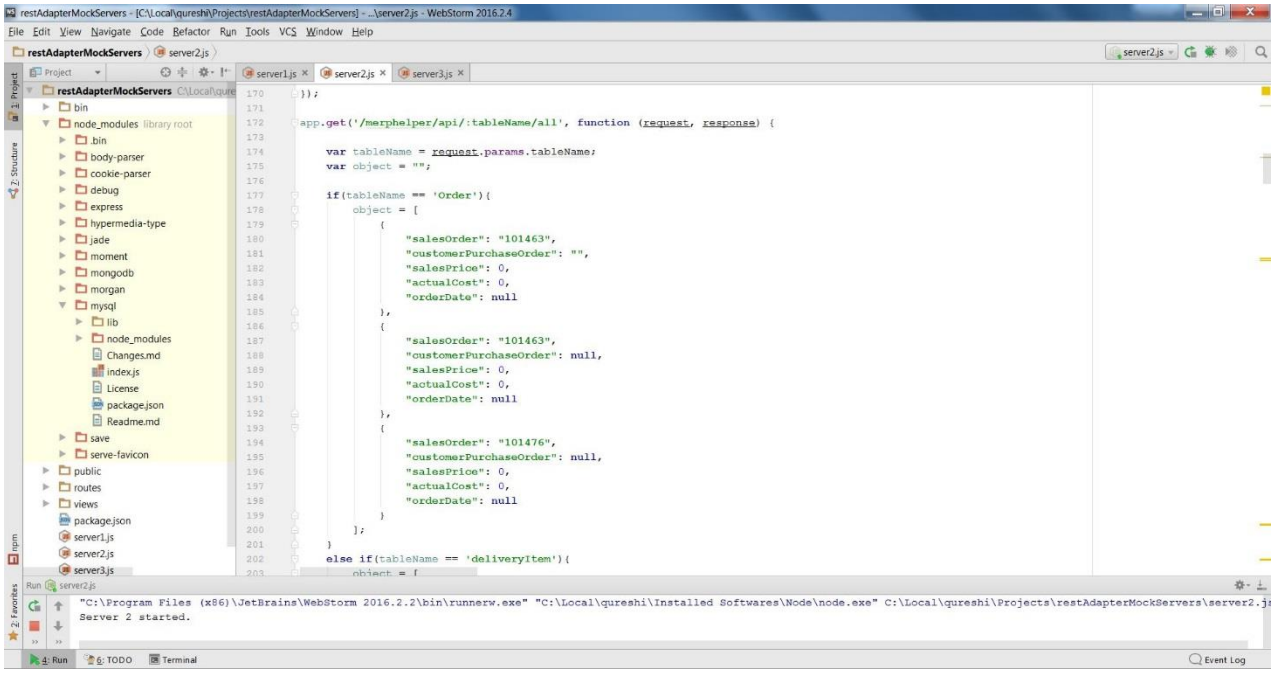


Figure 17. *Mockup server for the REST adapter data*

In this implementation, the user uses the RM interface to communicate with the LSH. For the REST adapter, the user needs to input the id of the hub, the function block id, the URL and method of the REST source. The REST adapter tab for the RM is shown in the below figure.

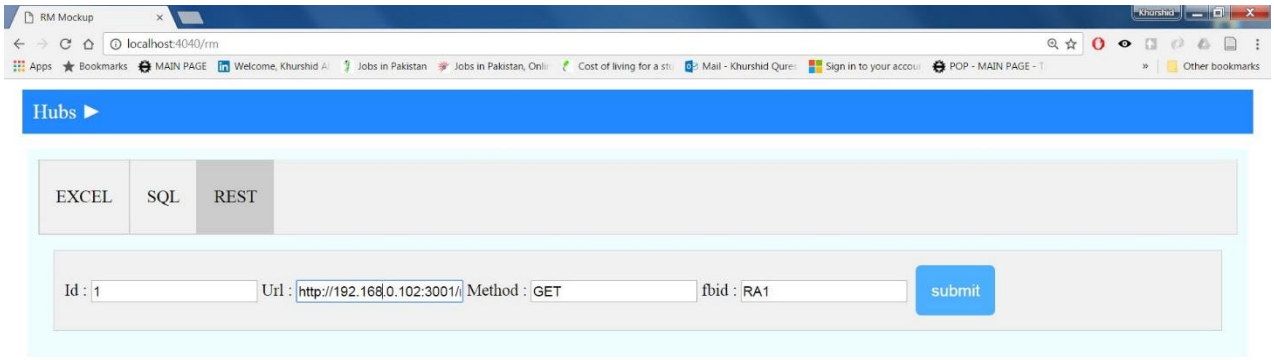


Figure 18. *Resource Manager user interface for sending half configuration to the REST adapter*

RM converts these fields and puts them in a JSON object. This JSON object is then forwarded to the PubSub module. Here, the PubSub module converts this JSON object into the configuration JSON object. This configuration is represented in the figure below:

```
{
    "configuration": {
        "sources": {
            "id":"1",
            "url":"http://130.230.181.107:3001/merphelper/api/database/Order",
            "method":"GET",
            "fbpid":"RA1",
            "fbid":"pid_eu.plant.restAdapter"
        }
    }
}
```

Figure 19. *Input JSON configuration for the REST adapter*

This configuration, when received in the REST adapter instance, initiates a Source Manager (SM) class object and uses it to store these configuration details. After that, and Engine class object is initiated that runs the adapter logic to fetch the column name data from the REST source and uses

the View Generator class object to create an HTML string which it sends back to the RM using the PubSub Module. Upon receiving the HTML, the web browser renders it and converts it into a table representing the column name data. The below figure is a screenshot showing the render table in the RM user interface.

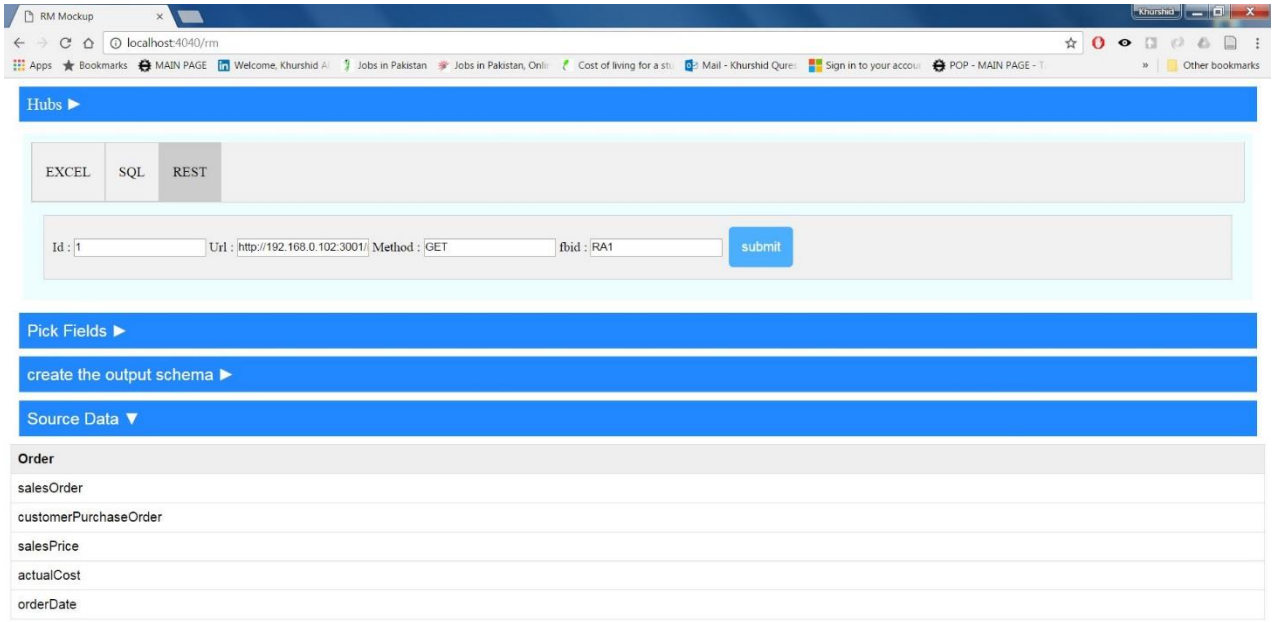


Figure 20. *Resource Manager showing fetched column fields from a REST data source on its user interface*

The user can then select the fields from these for which he needs the column data. The selected fields appear in the *Pick Field* as represented in the figure below. The *Send Fields* button can then be used to send the selected fields to the FBI in the form of an array.

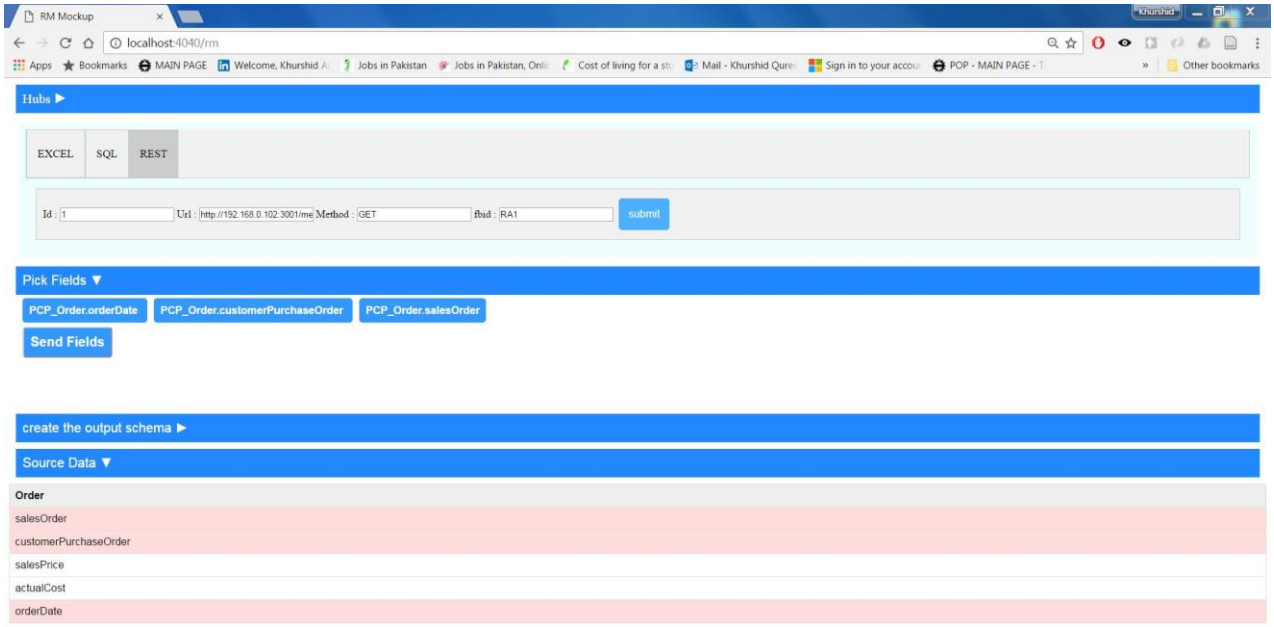


Figure 21. *Resource Manager after selecting the fields we need to fetch the data for from REST data source*

The *onMessageReceived* function of REST adapter instance handles the input message from the RM according to the *JMSType* fields included in the headers of the message. In this case, the *JMSType* of this message is set to "getData" by the PubSub module. The figure below shows the *onMessageReceived* function of the REST adapter. The code checks it for three different *JMSTypes*, "changeConfiguration", "getHTML" and "getData" and behaves accordingly.

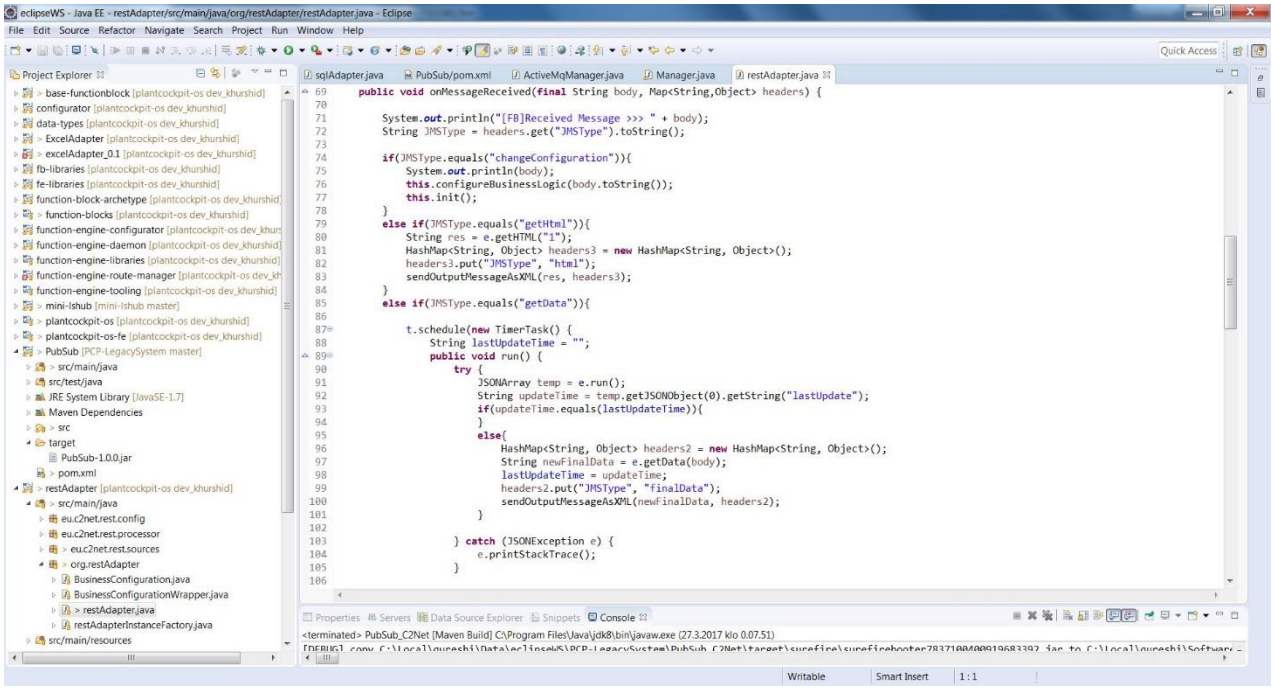


Figure 22. *onMessageReceived function of REST adapter to handle data according to the input JMSType*

After receiving the request for data, the Adapter fetches the data of the entire source and parses it to get the fields required by the user. To get all the data, the REST adapter uses the Run HTTP Request class object and parses the data and use the parsing functionality of the *getData* function of the Engine class object to parse it. The parsed data is put into a JSON object and is returned to the PubSub module, which forwards it the RM. RM, also exposes the acquired data on a HTTP endpoint so the user can view it in a web browser. The web browser view of data can be seen in the figure below.

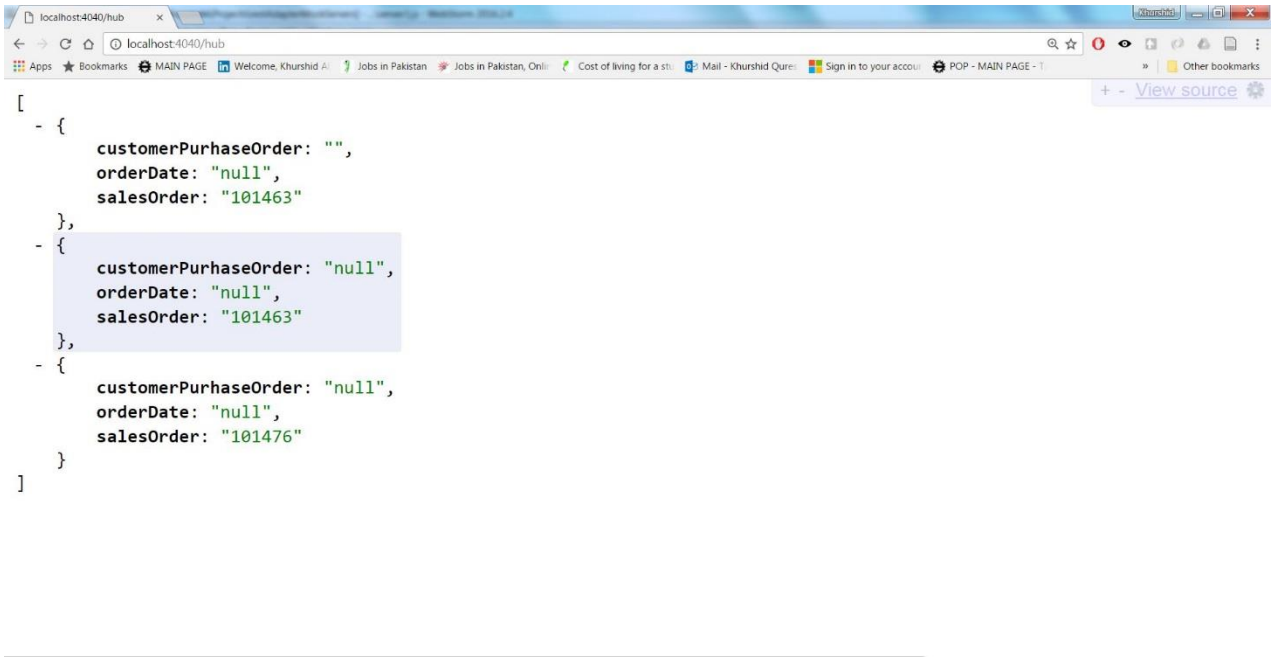


Figure 23. *Web browser showing the final data fetched from Rest data source in the form of a JSON object*

After this, until the instance of this particular Adapter is destroyed by the user, it checks its data source for any updated data every five seconds. If there is any change in the data from the previously checked value, it fetches the data again and updates the resource manager.

The next use case discussed in this section is the SQL adapter. The SQL adapter also follows the same information flow as that of the REST adapter. Details of SQL adapter are discussed in the next heading.

4.3.2 SQL Adapter

The purpose of SQL adapter is to fetch data from SQL database. An SQL database stores data in form of tables with relations to each other. This type of database architecture is known as relational database. In this implementation, only one particular SQL database server is realized known as MySQL. The snapshot below shows a view of MySQL Workbench used to create and manage mock up database for the implementation and testing of SQL adapter.

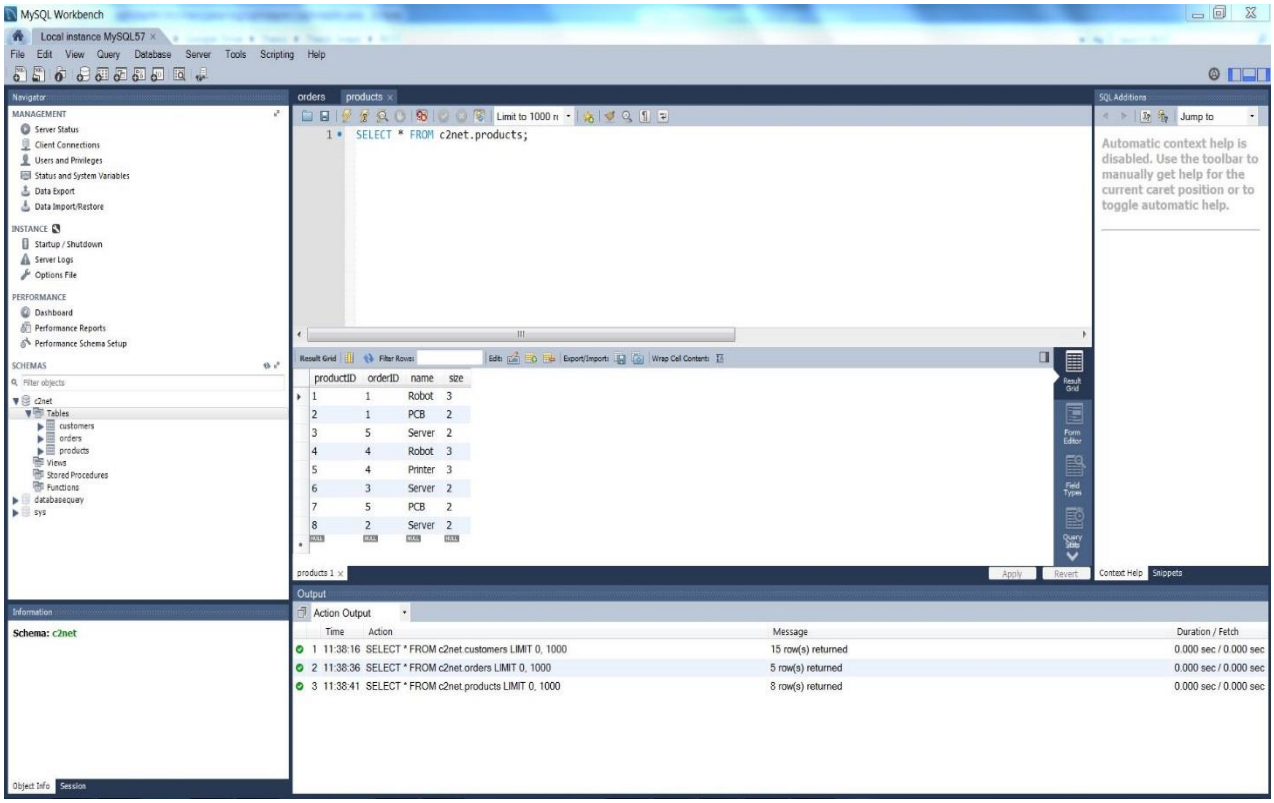


Figure 24. *Mock Database to work with SQL adapter*

Like the REST adapter, SQL adapter also needs configuration fields from the user to create its instance. User can use the RM user interface to send the half configuration to the LSH to configure the SQL adapter instance. The figure below illustrates the web based RM mock up interface used to send the configuration to the SQL adapter.

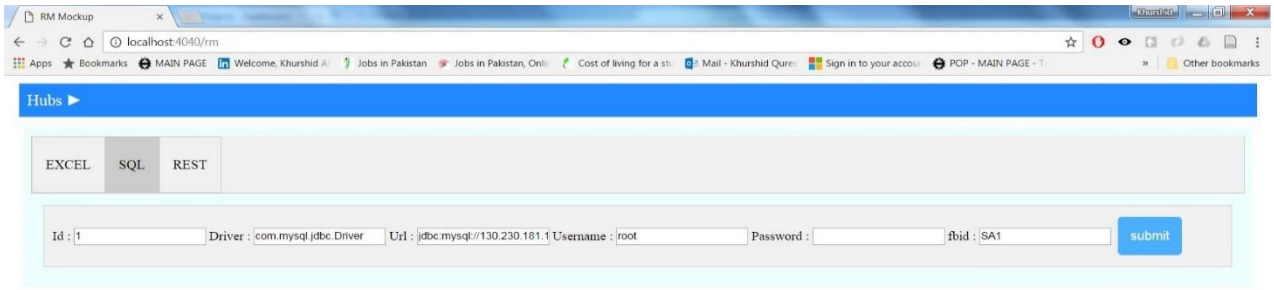


Figure 25. *Resource Manager user interface for sending half configuration to the SQL adapter*

The fields from RM are converted into JSON object and finally inserted into an XML to complete the configuration generation in the PubSub module. The JSON configuration for the SQL adapter contains the driver details for the database, the URL hosting the database, the credentials for database validation and the id of the LSH. Complete JSON configuration can be visualized by the figure below.


```
{
    "configuration": {
        "sources": {
            "id": "1",
            "driver": "com.mysql.jdbc.Driver",
            "url": "jdbc:mysql://130.230.181.123:3306/databasequery?useSSL=false",
            "credentials": {
                "username": "some",
                "password": "1234"
            },
            "fbpid": "RA1",
            "fbid": "pid_eu.plant.restAdapter"
        }
    }
}
```


Figure 26. *Input JSON configuration for the SQL adapter*

Similarly, the REST adapter, like SQL adapter also has the same FBI structure. After receiving the configuration, it uses the *Engine*, *Source Manager* and the *View Generator* class objects to fetch

and convert the column name fields and change them in HTML string. The HTML string is then sent to the RM, where the web browser renders the HTML to form an interactive table for the user. The screenshot of the web browser with the fetched fields can be seen in the figure below.

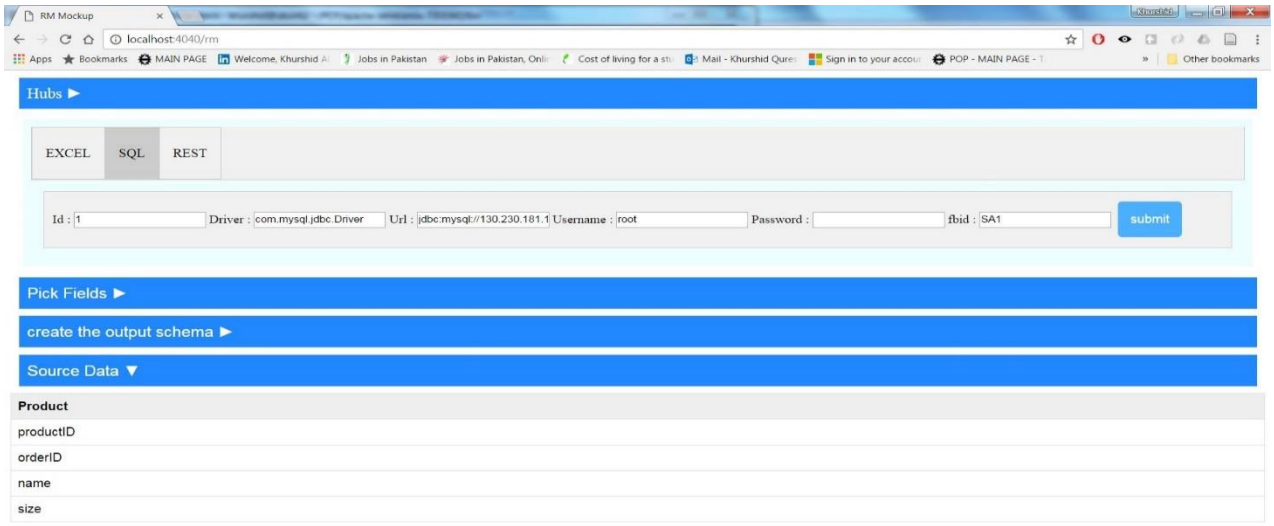


Figure 27. *Resource Manager showing fetched column fields from a SQL database on its user interface*

The user can select the required column names from the interface. As the user clicks on a field, it appears in the *Pick Field* tab. The *Send Fields* button is then pressed to send those fields to the LSH. The web browser screenshot below shows the view with selected fields.

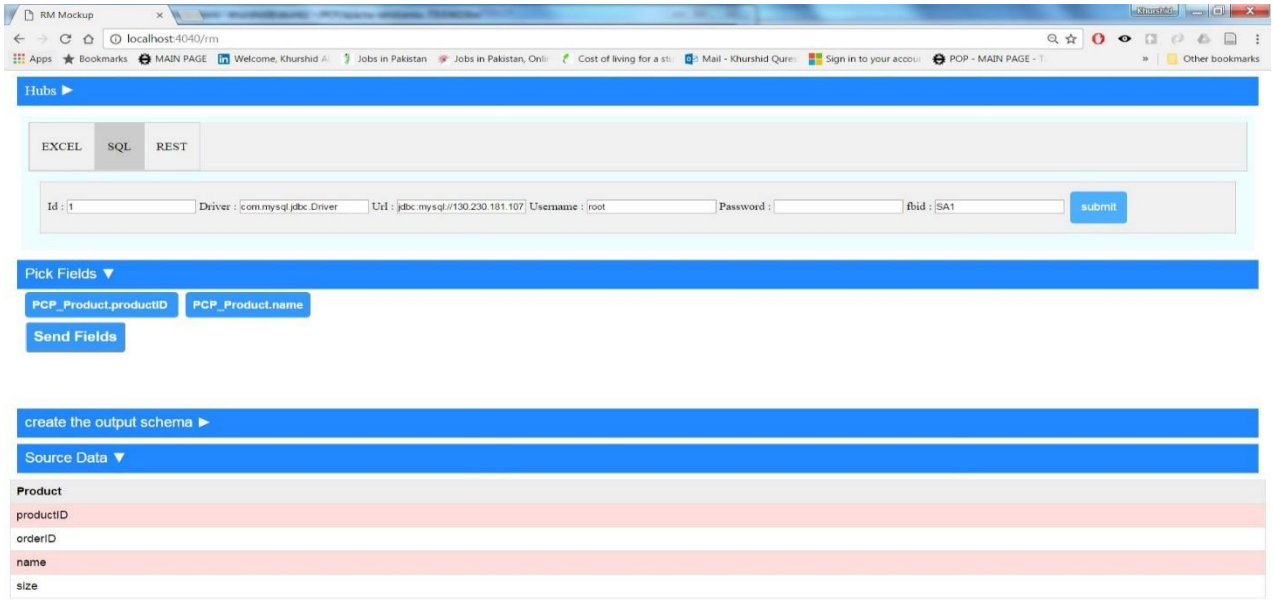


Figure 28. *Resource Manager after selecting the fields we need to fetch the data for from SQL database*

The PubSub module receives the above fields from the RM and forwards it to the FBI. The *on-MessageReceived* function of the SQL Adapter has a similar structure to check the input for three different cases, i) changeConfiguration, i) getHTML, iii) getData. This time the getData check will

run to fetch the column data from the data source. The structure of the onMessageReceived function can be seen in the figure below.

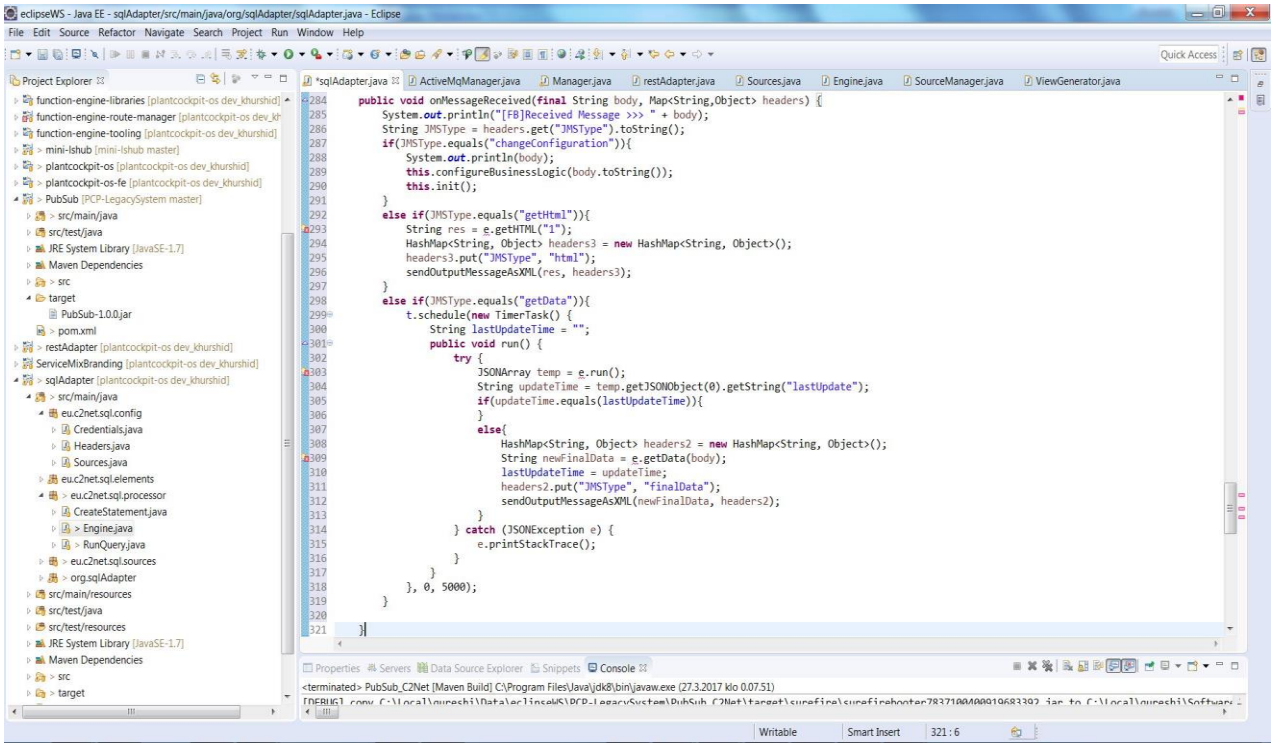


Figure 29. *onMessageReceived function of SQL adapter to handle data according to the input JMSType*

After receiving the selected column name fields in the *onMessageReceived* function, the *getData* function logic uses the *Data Fetching* class object to get all the data of that particular configured source. It then parses the data according to user requirements and sends back only the fields, which were selected by the user. The data is delivered to the RM through the output topic of the FBI through the PubSub module. After getting the data, the RM exposes the received data to an endpoint, which the user can use to attain it or check it in a web browser. The screenshot below illustrates an example of the type of data that can be received using the SQL adapter.

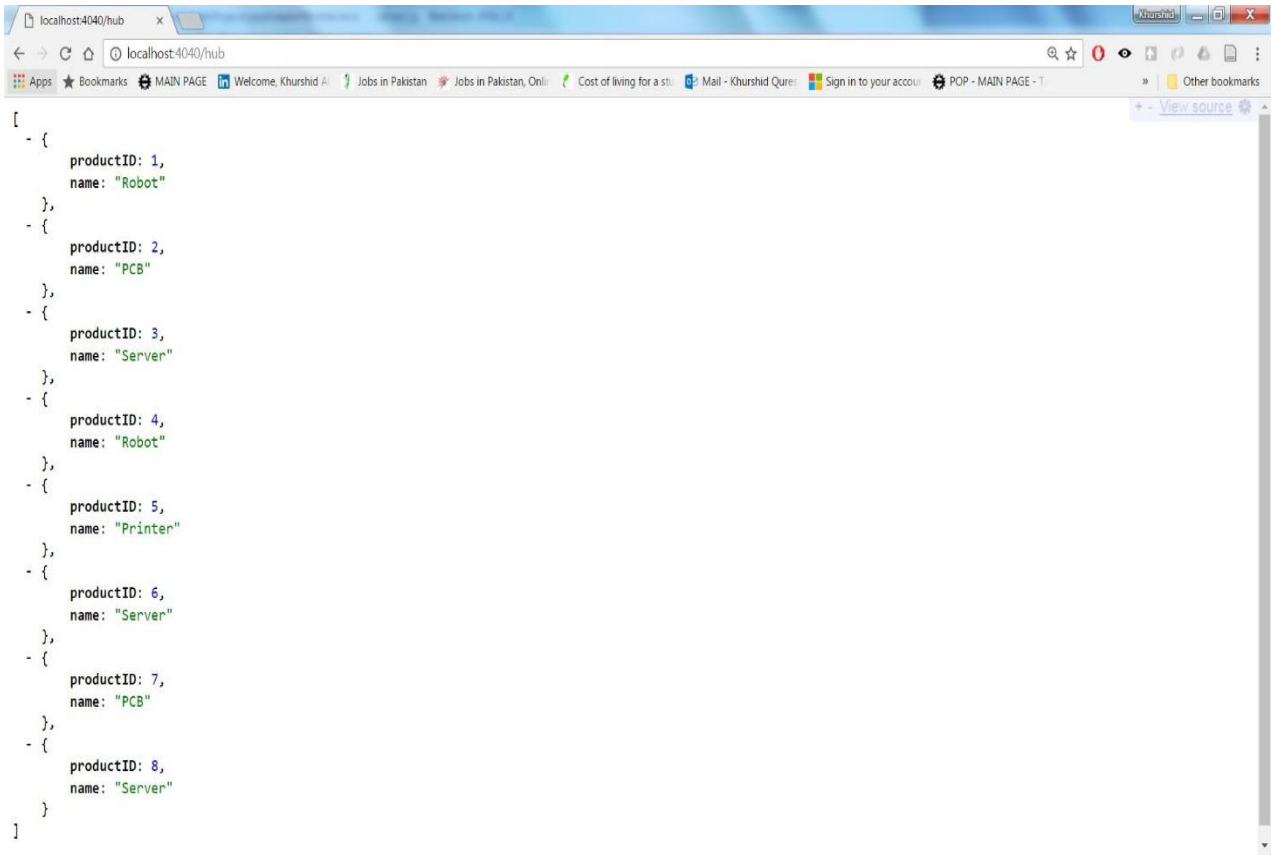


Figure 30. *Web browser showing the final data fetched from SQL database in the form of a JSON object*

4.3.3 Excel Adapter

Excel adapter is implementation of data adapter used to fetch data from .xls files. The architecture of the implementation is identical to that of the above two discussed use cases. In case of Excel adapter to work, the target Excel file needs to be hosted by an FTP server. The adapter needs the credentials of the server and the name of the particular Excel file to access its data. The below figure shows the configuration JSON needed by the LSH to create and configure the Excel adapter.


```
{
    "configuration": {
        "source": {
            "path": "/Examplesheets",
            "hostname": "192.168.1.176",
            "password": "",
            "name": "P1.xls",
            "id": "1",
            "type": "Excel",
            "username": "C2NET"
        }
    }
}
```


Figure 31. *Input configuration for Excel Adapter*

The user again needs to send the configuration with the help of RM user interface. The required fields are the hub id, path, name of Excel file, and credentials of the FTP server. The RM user interface can be visualized by the figure bellow.

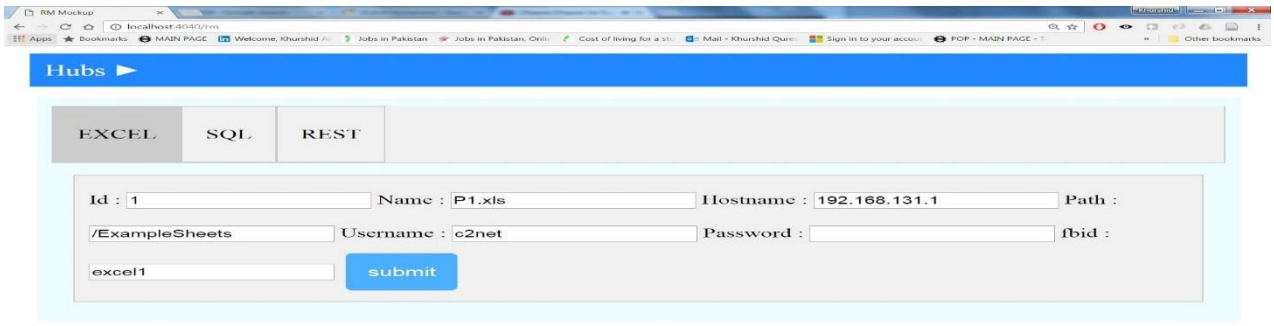
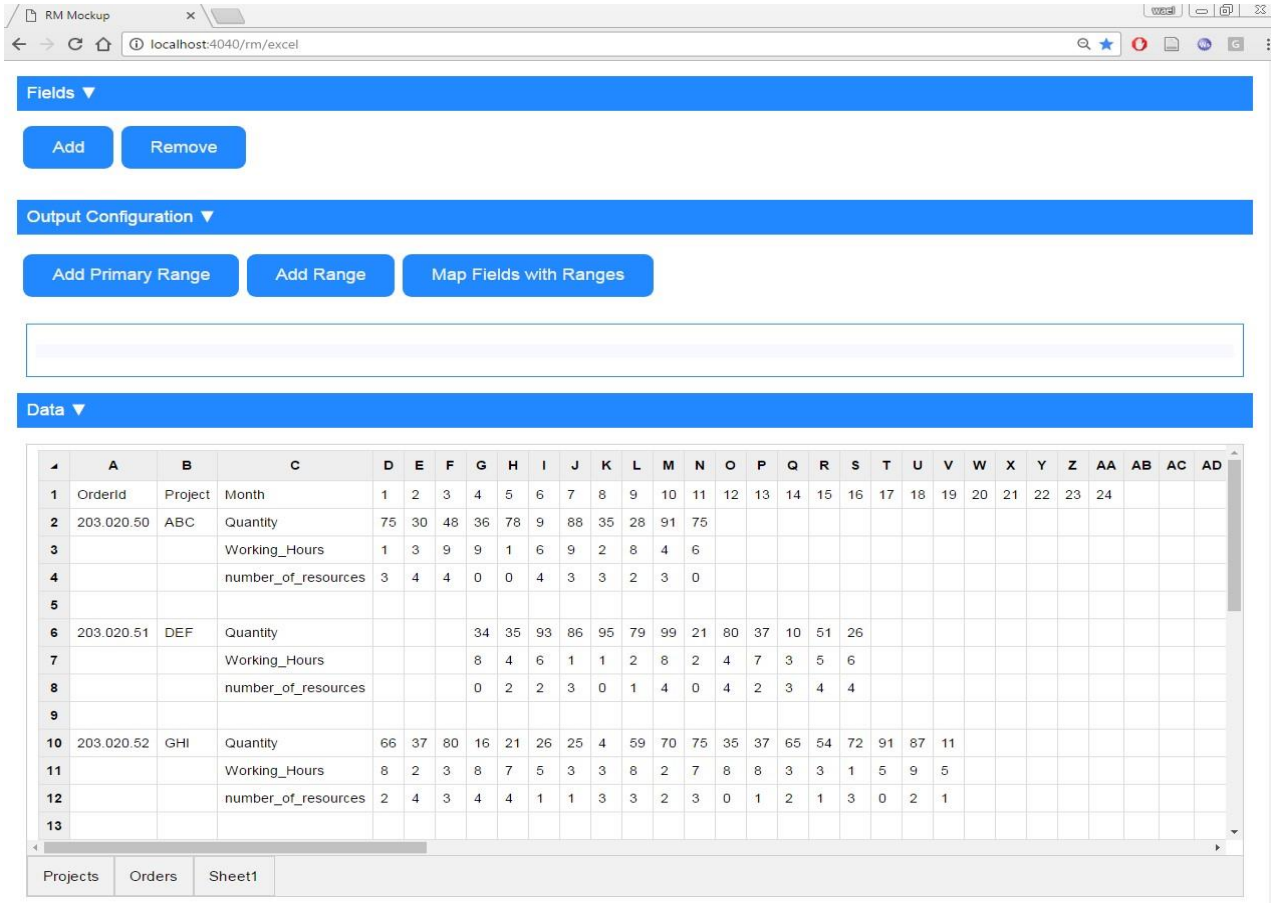


Figure 32. *RM interface for the Excel adapter*

The adapter after getting the initial configuration configures the FBI and uses the *Engine* object class to manage its resources through the *Source Manager* class object. The *Engine* fetches the column name data from the data sources similarly to the previous two adapters. The data is converted into an HTML string and sent to the RM via the PubSub where the web browser renders the HTML and shows it to the user. The screenshot below shows the returned column name data from the data source in a table format. The user can then select the fields according to his requirements. The selected fields will appear in the Fields tab. The user interface shows some additional features for configuration of the Excel adapter, which can be used to fetch Excel file data by manually entering the cell ranges to acquire the data according to his needs.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD
1	OrderId	Project	Month	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24			
2	203.020.50	ABC	Quantity	75	30	48	36	78	9	88	35	28	91	75																
3			Working_Hours	1	3	9	9	1	6	9	2	8	4	6																
4			number_of_resources	3	4	4	0	0	4	3	3	2	3	0																
5																														
6	203.020.51	DEF	Quantity				34	35	93	86	95	79	99	21	80	37	10	51	26											
7			Working_Hours				8	4	6	1	1	2	8	2	4	7	3	5	6											
8			number_of_resources				0	2	2	3	0	1	4	0	4	2	3	4	4											
9																														
10	203.020.52	GHI	Quantity	66	37	80	16	21	26	25	4	59	70	75	35	37	65	54	72	91	87	11								
11			Working_Hours	8	2	3	8	7	5	3	3	8	2	7	8	8	3	3	1	5	9	5								
12			number_of_resources	2	4	3	4	4	1	1	3	3	2	3	0	1	2	1	3	0	2	1								
13																														

Figure 33. *RM with fetched column name fields from the data source*

The selected fields are sent to the FBI via the PubSub module. The FBI uses the requirements from the message to fetch the data from the data source. It then parses and prepares the data in the final format required by the user. In the case of Excel adapter, the format of the output data is similar to the output data acquired from the SQL and REST adapters, as in both of these adapters the final data is in form of JSON object pairs. The final data can be visualized through the below figure.

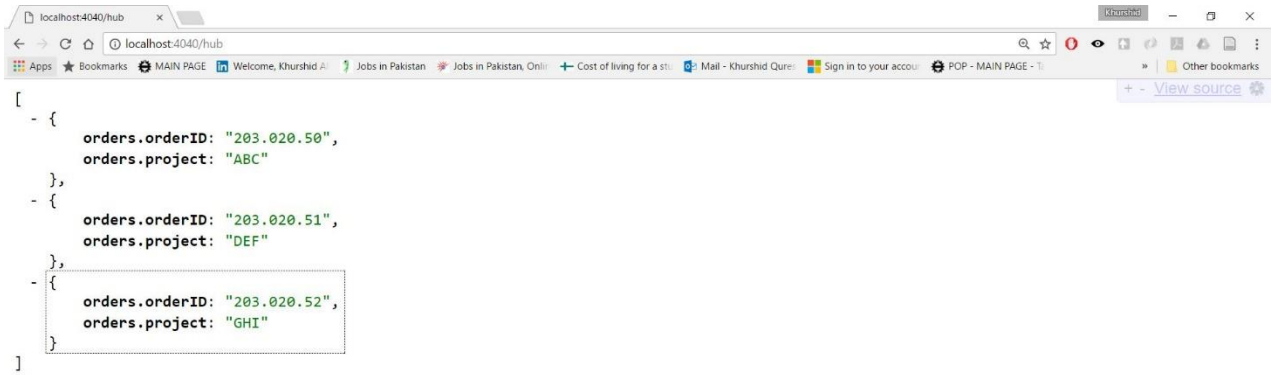


Figure 34. *Final data output of the Excel adapter*

4.4 Comparison with Legacy Systems

Figures 35 and 36 show an example Excel sheet and its corresponding JSON object, respectively. When performing this conversion from the Excel sheet to the JSON object manually, the user has to go through the Excel sheet row by row, and write the corresponding cell values one by one. For instance, the red box in Figure 35 translates into the red box in Figure 36; likewise for the blue box. This method of manual data entry can be quite error prone, especially when there is a significant amount of data, and the data is complex, e.g. floating point numbers (Final Percentage) and non-traditional data types (Final Letter Grade).

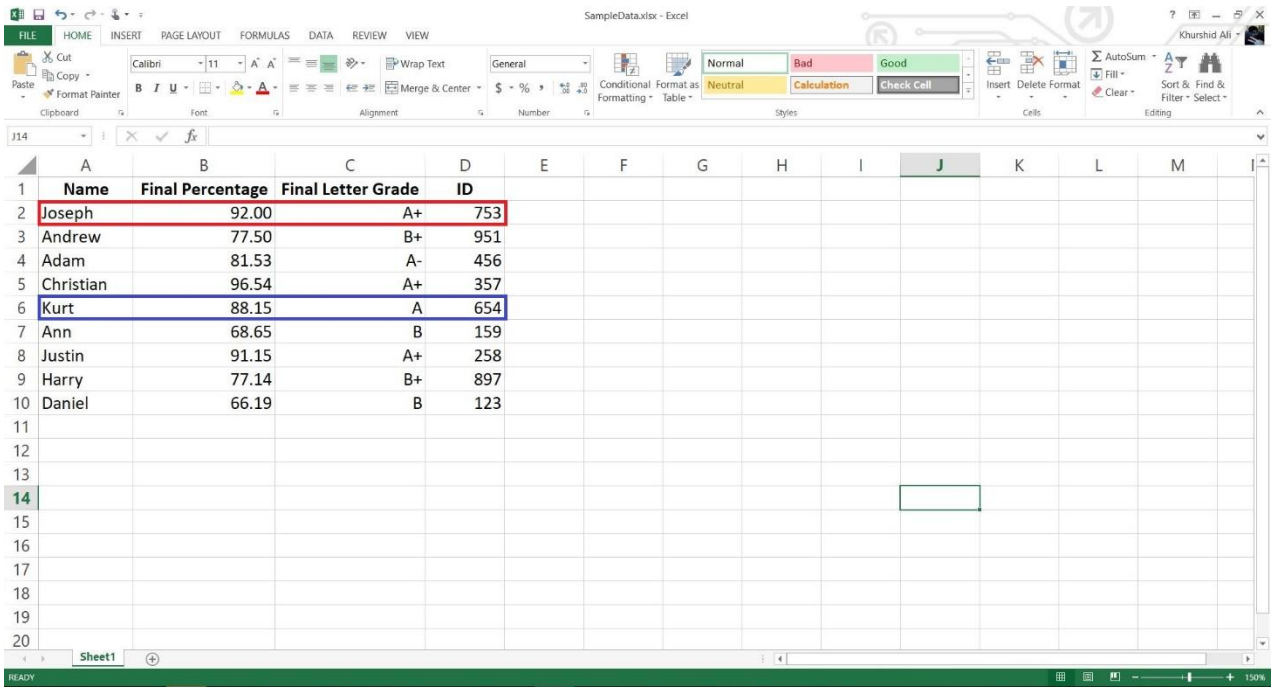


Figure 35. *Example Excel sheet for manual data migration.*

```
[
    {
        "Name": "Joseph",
        "Final Percentage": "92.00",
        "Final Letter Grade": "A+",
        "ID": "753"
    },
    {
        "Name": "Andrew",
        "Final Percentage": "77.50",
        "Final Letter Grade": "B+",
        "ID": "951"
    },
    {
        "Name": "Adam",
        "Final Percentage": "81.53",
        "Final Letter Grade": "A-",
        "ID": "456"
    },
    {
        "Name": "Christian",
        "Final Percentage": "96.54",
        "Final Letter Grade": "A+",
        "ID": "357"
    },
    {
        "Name": "Kurt",
        "Final Percentage": "88.15",
        "Final Letter Grade": "A",
        "ID": "654"
    },
    {
        "Name": "Ann",
        "Final Percentage": "68.65",
        "Final Letter Grade": "B",
        "ID": "159"
    },
    {
        "Name": "Justin",
        "Final Percentage": "91.15",
        "Final Letter Grade": "A+",
        "ID": "258"
    },
    {
        "Name": "Harry",
        "Final Percentage": "77.14",
        "Final Letter Grade": "B+",
        "ID": "897"
    },
    {
        "Name": "Daniel",
        "Final Percentage": "66.19",
        "Final Letter Grade": "B",
        "ID": "123"
    },
]
```


Figure 36. *Manually created JSON data from the example Excel Sheet.*

On the other hand, when performing this task with LSH, the amount of user effort is drastically reduced. All the user needs to do is 1) enter the configuration, 2) select the desired column names, and 3) get the ready JSON data from an http endpoint. This completely eliminates manual data acquisition, saving time and reducing the probability of error.

5. RESULT AND ANALYSIS

In this Chapter the results of the research and analysis have been presented in view of the problem defined in Chapter 1. The findings discovered during the course of the study along with the research constraints exposed have been elaborated in this Chapter. It is the integral section of the thesis, which merges the theory studies in Chapter 2 with the practical implementation done in Chapter 4.

5.1 Overview of Problem

The theory discussed in the Chapter 2 of the thesis has covered almost all the aspects of the issue. The challenges of legacy systems and data quality have been the basis of whole study done. The theory revealed the importance of data in the organizations and the problem of its timely acquisition according to user needs.

The importance of data generation in the companies especially in the supply chain cannot be overlooked. From order placement to collection of raw material, from product manufacturing to product delivery, data is produced in each stage of the supply chain. Mostly organizations have deployed ERP systems for the effective categorization of this data. However, to retrieve this ERP data from the legacy systems according to user provided format in a timely yet efficient way was somehow lagging.

5.2 Revisiting Research Questions

The research questions, which have been highlighted in the Chapter 3 of the thesis, have been probed throughout the Chapter 4. In this section, an analysis of the solution of thesis questions has been presented.

The first question was:

I. How data acquisition adapters can mitigate the challenges of coping with legacy systems?

The research reveals that the challenges of legacy systems have been addressed by the development of adapters to fetch the ERP data from heterogeneous legacy systems. As the issues of legacy systems discussed in section 2.2 consist of their rigid structure, resistance to modification and large source codes that impedes their upgradation; the data collection framework designed in this thesis has somehow satisfied almost all the previously mentioned problems.

Firstly, the legacy systems need a layer of interoperability, which can support their compatibility issue and keep their data accessible according to market demands. This layer of compatibility has been developed in this thesis with the help of cloud computing technology. This layer is the real time data collection framework, which has been remodified according to the user needs. Thus, it enhances the compatibility of legacy systems so that their seamless integration can be ensured.

Secondly, the legacy systems are kept readily available according to the business needs with the help of this information-sharing layer. It allows the access of data from heterogeneous data sources (legacy systems), transfers it to the LSH for data harmonization and management, and then forwards it to the user, which in this case is C2NET.

Lastly, this data acquisition framework allows the adaptability of legacy systems. This means that any changes or updates, which are required for the upgradation of legacy systems, can be done without affecting the already stored data of the legacy systems. In this way, legacy systems are able to adapt to changes in the environment in a time efficient approach.

To conclude, the first research question has been aptly answered during the course of the thesis. Now, the next question was:

II. How information flow can be enhanced to retrieve the readily available data required by user?

As discussed in Chapter 2, the supply chain networks are mostly dependent on complex information systems, which are unable to give them comprehensive view of the supply chain. This goal has been achieved through C2NET platform, which is the basis of designing a framework of this thesis.

The approach of this research talks proposes a generic platform for the information delivery. The significant feature of this DCF is to provide a single functional channel for the flow of information. Regardless of heterogeneous data types or diverse sources, the designed mainframe accomplishes the objective of providing a user desired information gathering method.

The exchange of information has been improved to facilitate the production and management systems. The collaborative approach used to design the LSH of DCF has made the optimization of resources better. This in short has increased the efficiency of business.

Moreover, the flow of information between the supplier and user has been improved. Similarly, PubSub used in the framework provides loose coupling to enhance the information flow between client and server. Furthermore, it has better scalability as compared to any traditional client server information-sharing network.

5.3 Findings and Framework

This section will enlighten about the results and findings, which are derived from the implementation of the designed framework. The statistical analysis of the problems and the solutions acquired in light of the implementation is discussed below.

- **Time Efficiency**
  As discussed in Chapter 2 of the thesis, the issue was ‘manual collection and recording of data used to take 30 to 50 percent of field supervisors’ time’ [19].

  The results of the implementation have found solution of this problem. To understand the methodology, assumption of data has to be taken. For instance, if there are ten rows of five

different fields in an excel file, the time taken to convert this data into JSON format will be one hour. Instead of manual data collection, the data will be collected through DCF of C2NET project. In this way, the user just needs to give the details of particular excel file to configure the LSH. Once the Legacy System Hub return the column name values in some milliseconds, the user will further take a couple of minutes to select the required columns. This is then submitted to LSH to acquire data. LSH provides the user with the furnished data in a JSON object as the user required in just a total of approximately five minutes.
The same is the case with SQL adapter in which the user does not have to write complex queries and just need to put configuration in a non-technical way to fetch the data required again in the desired JSON format. This signifies that the user is able to fetch the required data in a time efficient way with the help of implemented framework.

- **Reduction of Faulty Data**
  According to one estimate, there is 25 to 30 percent faulty data inclusive of missing or repetitive data [60].
  This issue has been addressed in the course of the research in the most suitable way. It is necessary to analyze the types of errors in data, which results due to manual data collection. The first is missing any data entry due to human error and the second is the chances of repeating the data value during the data entry task. Lastly, the inaccuracy of data can be caused by unrecognizing any repetitive data in the data source. All of these mistakes will lead to faulty, incorrect and inconsistent data.
  This research work has tried to mitigate the problem by providing a DCF platform where physical data entry tasks are not required. Firstly, the values are picked from the file automatically. Secondly, the system itself identifies any repetitive information. Hence, the JSON format diminishes the chance of any repeated entry in the data.

- **Maintenance Of Legacy System**
  As revealed in Section 2.2 ‘The cost of maintaining, operating and supervising the legacy systems is too high. It is so expensive that a survey analysis found that almost 85 to 90 percent of the company’s total budget is spent on the maintenance of legacy system [44]’.
  Regarding the cost of maintenance of legacy systems, the Function Block based approach, which is used in this thesis, will provide a better remedy. The adapters connected with LSH and C2NET have mitigated the trouble of upgradation of entire legacy systems. Keeping the technology advancement with time in mind, the system just requires the adapters to be upgraded. The companies do not need to worry about migrating or maintaining huge legacy system data. This will cut the cost of maintaining legacy systems completely or make it minimal.

The first two points are illustrated using an example in Section 4.4. The third point is discussed in Section 5.4.

6. CONCLUSION

This is the final Chapter of the thesis, which reiterates the accomplishment of the thesis in the light of the overall summary of the thesis. It also presents approach used for the validation of the research done. Finally, it states the recommendations resulted during the course of the study for future work.

6.1 Summary

From collection of raw materials to production and delivery of products or services, a huge amount of data is generated in supply chain. Enterprise Resource Planning (ERP) software has streamlined the supply chain management, by allowing inventory optimization through collection of data. Therefore, the availability of accurate and readily accessible data is required for the efficiency of supply chain. In this regard, a new wave of technology, Cloud based computing or often known as on demand computing of data and information has further supplemented the supply chain without the use of any hardware or software. This paper presented a mainframe of cloud-based collaboration of systems (C2NET) which has provided a platform for companies to enrich their supply chain. It has widely focused on the methodology given by C2NET to retrieve data from ERP systems of supply chain.

The research presented in this document has conveyed the deployment of function block based approach by working on three highly relevant use cases, which are REST, SQL and Excel. Legacy System Hub (LSH), which is a part of Data Collection Framework (DCF) used in C2NET Project, has been the main focal of this research. LSH has been built to provide a platform for the integration of ERP data by adapters with the C2NET.

6.2 Validation of Research

This research work has been done on the mainframe of C2NET project, which has received its funding from European Union's Horizon 2020 research and innovation program under grant agreement n° 636909. The functional validation of this research will be authorized after the examination of the implementation of the pilot project of C2NET. This validation will be conducted by analyzing the business performance where legacy systems are functional through questionnaire and results. In this way, the evaluation of real time JSON format information gathering and sharing will be supported by optimization and management of supply chain resources through cloud based platform.

As mentioned in Chapter 3 of this thesis the credibility of the research has been explored by doing quantitative and qualitative analysis. In addition, the active communication and guidance of the supervisors of this research work has mitigated the bias of the results. In conclusion, each step taken in the research work of this thesis has been done thoroughly and carefully to ensure its reliability.

6.3 Recommendations for Future Research

This thesis has been successful to provide a user-friendly interface for the users of C2NET. It has also supplemented the research done so far on the implementation of C2NET project. As a result, it has opened various research questions for the researchers in this field. Few of which have been mentioned below.

This thesis has presented a framework in which three data adapters have been designed for collection of data from three different data sources REST, SQL and Excel. However, future research can work on the implementation of multiple data sources apart from these three for the acquisition of data.

The research done above has focused on the development of adapters to fetch the data in JSON format for the user. However, in future, researchers can work on the deployment of two adapters, which can be used serially if JSON data is not required by C2NET. It means, one data adapter transfers the data to another data adapter. The second adapter after converting into non-JSON format will forward it to C2NET platform.

In addition to the above mentioned research arenas, the research limitations also provide with future work directions. For instance, the ability to work with diverse data sources other than SQL, REST and Excel can be explored in future.

Similarly, the generation of a new generic platform other than RM for the communication with LSH can be investigated in future. Lastly, the flexibility of adapters to work with other data sources can be examined in consequent research works.

In short, this research work has the potential to become the intriguing point for future developments in design and development of data collection framework to get the required information in user-desired format.

7. REFERENCES

[1] Wand, Y. and Wang, R.Y. (1996), “Anchoring data quality dimensions in ontological foundations”, Communications of the ACM, Vol. 39 No. 11, pp. 86-95.

[2] Pipino, L.L., Lee, Y.W. and Wang, R.Y. (2002), “Data quality assessment”, Communications of the ACM, Vol. 45 No. 4, pp. 211-18.

[3] Kanaracus, C. 2014. Cost of troubled SAP project will skyrocket to nearly $1 billion, audit says [e-zine]. Computerworld. Published on 03.10.2014 [accessed on 02.10.2016]. Available: http://www.computerworld.com/article/2691661/cost-of-troubled-sap-project-will-skyrocket-to-nearly-1-billion-audit-says.html.

[4] Levitin, A.V. and Redman, T. (1998), “Data as a resource: properties, implications, and prescriptions”, Sloan Management Review, Vol. 40 No. 1, pp. 89-101.

[5] Friedman, T., Feinberg, D., Beyer, M.A., Gassman, B., Bitterer, A., Newman, D., Radcliffe, J., White, A., Paquet, R., DiCenzo, C., Logan, D., Blechar, M., Knox, R.E., Bell, T. and Shegda, K.M. (2006), Hype Cycle for Data Management, 2006, GartnerGroup Research, July 6, ID#G00140057.

[6] Kahn, B., Strong, D. and Wang, R. (2003), “Information quality benchmarks: product and service performance”, Communications of the ACM, Vol. 45 No. 4, pp. 184-92.

[7] Ballou, D. and Pazer, H. (1985), “Modeling data and process quality in multi-input multi-output information systems”, Management Science, Vol. 31 No. 2, pp. 150-62.

[8] Madnick, S., Wang, R. and Xian, X. (2004), “The design and implementation of a corporate householding knowledge processor to improve data quality”, Journal of Management Information Systems, Vol. 20 No. 1, pp. 41-9

[9] Tayi, G.K. and Ballou, D.P. (1998), “Examining data quality”, Communications of the ACM, Vol. 41 No. 2, pp. 54-7.

[10] Cappiello, C., Francalanci, C. and Pernici, B. (2003), “Time-related factors of data quality in multi-channel information systems”, Journal of Management Information Systems, Vol. 20 No. 3, pp. 71-91.

[11] Lederman, R., Shanks, G. and Gibbs, M.R. (2003), “Meeting privacy obligations: the implications for information systems development”, Proceedings of the 11th European Conference on Information Systems (ECIS), Naples, Italy, 16-21 June, available at: http://is2.lse.ac.uk/asp/aspecis/20030081.pdf (accessed 29 June 2009).

[12] Watts, S., Shankaranarayanan, G. and Even, A. (2009), “Data quality assessment in context: a cognitive perspective”, Decision Support Systems, Vol. 48 No. 1, pp. 202-11.

[13] Marsh, R. (2005), “Drowning in dirty data? It’s time to sink or swim: a four-stage methodology for total data quality management”, Database Marketing & Customer Strategy Management, Vol. 12 No. 2, pp. 105-12.

[14] ISA-95.com. Technology ISA-95. [WWW]. [Accessed on 2.10.2016]. Available at:

http://www.isa-95.com/subpages/technology/isa-95.php

[15] ISA.com. The International Society of Automation. [WWW]. [Accessed on 8.10.2016]. Available at: http://www.isa.org

[16] Jean Pierre Lorré. 2015. 1st Data collection from resources virtualization: legacy systems integration Work package [ONLINE] Available at: http://c2net-project.eu/documents/10184/232801/D3.3_636909_1st+Data+collection+from+resources+virtualisation.pdf/6093934d-be4f-4614-8094-1766ddff344d. [Accessed 23 October 2016].

[17] Stouffer, Keith, Victoria Pillitteri, Suzanne Lightman, Marshall Abrams, and Adam Hahn. "Guide to Industrial Control Systems (ICS) Security." (2015): n. pag. Web. [Accessed 31 October 2016].

[18] I. Davidson, M. Skibniewski, Simulation of automated data collection in buildings, J. Comput. Civil Eng. 9 (1) (1995) 9-20.

[19] B. Mccullouch, Automating field data collection in construction organizations, in: Proceeding of the 1997 ASCE Construction Congress Mimleapolis, Mimlesota, 1997, pp. 957-963.

[20] G.S. Cheok, R.R. Lipman, C. Witzgall, J. Bernal, W.C. Stone, NIST Construction Automation Program Rep. No: 4, Non-Intrusive Scanning Technology for Construction Status Determination, Gaithersburg, Md, 2000. [25] R. Navon, Automated project performance control of construction projects, Autom. Constr. 14 (4) (2005) 467-476.

[21] S. Taneja, A. Akcamete, B. Akinci, J.H. Garrett, E.W. East, L. Soibelman, Analysis of Three Indoor Localization Technologies to Support Facility Management Field Activities, in: Proceeding of International Conference on Computing in Civil and Building Engineering, Nottingham, UK, 2010.

[22] Bennett, K. "Legacy Systems IEEE Software 12.1 (1995): 19-23. Web.: Coping with Success."

[23] Nelson H. Weiderman, John K. Bergey, Dennis B. Smith, and Scott R.Tilley, “Approaches to Legacy System Evolution,” Dec. 1997.

[24] W. Qifeng, “Semantic Framework Model-Based Intelligent InformationSystem Integration Mode for Manufacturing Enterprises,” in SecondInternational Symposium on Intelligent Information TechnologyApplication, 2008. IITA ’08, 2008, vol. 1, pp. 223–227.

[25] Y. Manolopoulos and Institute for Systems and Technologies ofInformation, Control and Communication, Eds., Enterprise information *systems: 8th International Conference, ICEIS 2006, Pahpos, Cyprus,May 23-27, 2006: revised selected papers*. Berlin: Springer, 2008.

[26] S. K. Makki, “The Integration and Interoperability Issues of Legacy andDistributed Systems,” in *Seventh International Conference on Web-AgeInformation Management Workshops, 2006. WAIM ’06*, 2006, pp. 21–21.

[27] Wu, B., Lawless, D., Bisbal, J., Grimson, J., Wade, V., O'Sullivan, D., & Richardson, R. (1997). Legacy Systems Migration: A Method and its Tool-Kit Framework. Proceedings of the APSEC'97/ICSC'97: Joint 1997 Asia Pacific Software Engineering Conference and International Computer Science Conference, Hong Kong, China, 312-320. doi:10.1109/APSEC.1997.640188

[28] Meng, F., Qu, Z., &Guo, X. (2013). Refactoring Model of Legacy Software in Smart Grid based on Cloned Codes Detection. IJCSI International Journal of Computer Science Issues, 10(1), 296-303.

[29] van Deursen, A., Klint, P., &Verhoef, C. (1999). Research issues in the renovation of legacy systems (pp. 1- 21). Springer Berlin Heidelberg.

[30] Rahgozar, M., &Oroumchian, F. (2002, October). A Transformational Approach for Legacy Systems' evolution. In Submitted for publication in: 2002 WSEAS International Conference on Applied Mathematicsand Computer Science (AMCOS'02), Copacabana, Rio De Janeiro.

[31] Sneed, H.M. (2006). Wrapping Legacy Software for Reuse in a SOA. Multikonfer-enzWirtschaftsinformatik, 2, 345-360.

[32] Stehle, E., Piles, B., Max-Sohmer, J., & Lynch, K. (2008). Migration of Legacy Software to Service Oriented Architecture. Department of Computer Science Drexel University Philadelphia, PA, 19104, 2-5.

[33] Khadka, R., Saeidi, A., Idu, A., Hage, J., & Jansen, S. (2013). Legacy to SOA Evolution: A Systematic Literature Review. In In AD Ionita, M. Litoiu, & G. Lewis (Eds.) Migrating Legacy Applications: Challenges in ServiceOriented Architecture and Cloud Computing Environments.

[34] Rahgozar, M., &Oroumchian, F. (2002). A practical approach for modernization of legacy systems. In EuroAsian Conference on Advances in Information and Communication Technology (ICT 2002)-Workshop onRecent progress in Computers and Communications.

[35] Nowakowski, W., Śmiałek, M., Ambroziewicz, A., Jarzębowski, N., &Straszak, T. (2012). Recovery and migration of application logic from legacy systems. Computer Science, 13(4), 53-70.

[36] National Association of State Chief Information Officers-NASCIO. (2008). Digital States at Risk: Modernizing Legacy Systems. Lexington: NASCIO publisher.

[37] Weiderman, N. H., Bergey, J. K., Smith, D. B., & Tilley, S. R. (1997). Approaches to Legacy System Evolution (No. CMU/SEI-97-TR-014). Carnegie-mellonunivpittsburgh pa software engi-neering inst.

[38] M. Yoshioka, T. Sudo, A. Yoshikawa, and K. Sakata, "Legacy system integration technology for legacy application utilization from distributedobject environment," Hitachi Rev., vol. 47, no. 6, pp. 284–290, 1998.

[39] C. Zhao, Q. Li, M. Wang, Y. Wang, and Y. Li, "An Agent Based Wrapper Mechanism Used in System Integration," in IEEEInternational Conference on e-Business Engineering, 2008. ICEBE '08,2008, pp. 637–640.

[40] R. C. Seacord, D. Plakosh, and G. A. Lewis, Modernizing Legacy Systems: Software Technologies, Engineering Processes, and BusinessPractices. Addison-Wesley Professional, 2003.

[41] A. M. Langer, "Legacy Systems and Integration," in Guide to Software Development, London: Springer London, 2011, pp. 179–212.

[42] H. Guo, G. Lu, Y. Wang, H. Li, and X. Chen, "RBAC-based access control integration framework for legacy system," in Web InformationSystems and Mining, Springer, 2010, pp. 194–201.

[43] Bisbal, J., Lawless, D., Bing, W., &Grimson, J. (1999). Legacy information systems: issues and directions. IEEE Software, 16(5), 103-111.

[44] Erlikh, L. (2000). Leveraging legacy system dollars for e-business. IT professional, 2(3), 17-23.

[46] Paradauskas, B., & Laurikaitis, A. (2006). Business knowledge extraction from legacy information systems, 35(3), 214-221.

[47] Gartner, Inc. (2007, December 20). The quest for talent: You ain't seen nothing yet. Gartner Research. Retrieved October, 7, 2017, from http://www.gartner.com/DisplayDocument?id=569115.

[48] van Geet, J., Ebraert, P., &Demeyer, D. (2010). Redocumentation of a Legacy Banking System. Proceedings of the Joint ERCIM Workshop on Software Evolution (EVOL) and International Workshop on Principles of Software Evolution (IWPSE), New York, USA, 33-41. doi: 10.1145/1862372.1862382.

[49] Lin, C. Y. (2008). Migrating to Relational Systems: Problems, Methods, and Strategies. Contemporary Management Research, 4(4), 369-380.

[50] Aversano, L., & Tortorella, M. (2004). An assessment strategy for identifying legacy system evolution requirements in eBusiness context. Journal of Software Maintenance and Evolution: Research and Practice, 16(4-5), 255-276.

[51] Brodie, M.L., &Stonebraker, M. (1995). Migrating Legacy Systems: Gateways, Interfaces and the Incremental Approach. Michigan: Morgan Kaufmann Publishers.

[52] Geetha, S. (2012). Possible Challenges of Developing Migration Projects. International Journal of Computers & Technology, 3(3), 463-465.

[53] Zhang, W., Berre, A. J., Roman, D., &Huru, H. A. (2009, October). Migrating legacy applications to the service Cloud. In 14th Conference companion on Object Oriented Programming Systems Languages and Applications (OOPSLA 2009) (pp. 59-68).

[54] Xia, W., & Lee, G. (2004). Grasping the complexity of IS development projects. Communications of the ACM, 47(5), 68-74.

[55] Khadka, R., Saeidi, A., Hage, J., & Jansen, S. (2013a). A Structured Legacy to SOA Migration Process and its Evaluation in Practice. Proceeding of the 7th MESOCA, Eindhoven, The Netherlands, IEEE.

[56] Khadka, R., Saeidi, A., Jansen, S., Hage, J., & Haas, G. P.(2013b). Migrating a Large Scale Legacy Application to SOA: Challenges and Lessons Learned. Proceedings of the 20th WCRE, Koblenz, Germany. IEEE.

[57] Seacord, R. C., Comella-Dorda, S., Lewis, G., Place, P., & Plakosh, D. (2001). Legacy System Modernization Strategies (No. CMU/SEI-2001-TR-025). CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST.

[58] Tonella, P., Torchiano, M., Du Bois, B., & Systä, T. (2007). Empirical studies in reverse engineering: state of the art and future trends. Empirical Software Engineering, 12(5), 551-571.

[59] van Geet, J. (2011). Reverse Engineering for Mainframe Enterprise Applications - Patterns and Experiences.

[60] Redman, T.C. 2008. Data driven: Profiting from your most important business asset. Boston, Massachusetts, Harvard Business Press. 12-29, 257 pp.

[61] Davenport, T.H., & Prusak, L. 1998. Working Knowledge: How Organizations Manage What They Know. Cambridge, Massachusetts, Harvard Business School Press. 224,4 pp.

[62] Zack, M.H. 1999. Managing codified knowledge. Sloan Management Review, Vol. 40(4), pp. 45-58.

[63] Haug, A., Arlbjørn, J.S., & Pedersen, A. 2009. A classification model of ERP system data quality. Industrial Management & Data Systems, Vol. 109(8), pp. 1053-1068.

[64] Haug, A., Zachariassen, F., & Van Liempd, D. 2011. The costs of poor data quality. Journal of Industrial Engineering and Management, Vol. 4(2), pp. 168-193.

[65] Wang, R.Y. & Strong, D.M. 1996. Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems, Vol. 12(4), pp. 5-34.

[66] McGilvray, D. 2008. Executing data quality projects ten steps to quality data and trusted information. Amsterdam, Morgan Kaufmann/Elsevier. 352 p.

[67] Olson, J 2003. Data Quality: The Accuracy Dimension. USA, Morgan Kaufmann Publishers. 300 p.

[68] Eckerson, W.W. 2002. Data quality and the bottom line. TDWI Report, The Data Warehouse Institute. 32 p.

[69] Redman, T.C. 2013. Data's credibility problem. Harvard Business Review, Vol. 91(12), pp. 84-88.

[70] Feldman, S. & Sherman, C. 2001. The high cost of not finding information. Framingham, Massachusetts, IDS. 10 p.

[71] Lundqvist, Magnus. "Information Demand and Use: Improving Information Flow within Small-scale Business Contexts." Diss. Linköping Studies in Science and Technology, 2007. Web.

[72] Virta, J. 2010. Application integration for production operations management usingOPC Unified Architecture. Master's Thesis. Espoo. Aalto University, School of Scienceand Technology. 59 p.

[73] Scholten, Bianca. The road to integration: a guide to applying the ISA-95 standard in manufacturing. Research Triangle Park NC: ISA, 2007. Print.

[74] Williams, T. 1998. Enterprise integration in the process industries. Presented at theWorldBatch Forum 1998.

[75] Jammes, F., and H. Smit. "Service-Oriented Paradigms in Industrial Automation." IEEE Transactions on Industrial Informatics 1.1 (2005): 62-70. Web.

[76] C. M. MacKenzie, K. Laskey, F. McCabe, P. F. Brown, and R. Metz, "Reference Model for Service Oriented Architecture V1.0. N.p., n.d. Web. 26 Jan. 2017.

[77] M. Schlütter, U. Epple, and T. Edelmann, "On Service-Orientation as a New Approach for Automation Environments",in ARGESIM report, vol. 35, Proceedings / MATHMOD 09, I. Troch, Ed, Vienna: ARGESIM Publ. House, 2009, pp. 2426–2431.

[78] DAMA. Data Management International. [WWW]. [Accessed on 29.1.2017]. Available at: http://www.dama.org

[79] IEC 61499-1/Ed.2: Function blocks - Part 1: Architecture. International ElectrotechnicalCommission, IEC, Nov. 2012. URL: http://www.iec.ch/.

[80] ValeriyVyatkin. „The IEC 61499 standard and its semantics". In: IndustrialElectronics Magazine, IEEE 3.4 (Dec. 2009), pp. 40–48. issn: 1932-4529. doi:10.1109/MIE.2009.934796.

[81] Thomas Strasser, AloisZoitl, James H. Christensen, and Christoph Sünder. „Designand Execution Issues in IEC 61499 Distributed Automation and Control Systems".In: Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEETransactions on 41.1 (Jan. 2011), pp. 41–51. issn: 1094-6977. doi: 10.1109/TSMCC.2010.2067210.

[82] Schmeling, Matthias. Effective Visualization of IEC 61499 Function Blocks: With the CAKeFEED Function Blocks Editor. Thesis. Kiel, n.d.N.p.: n.p., n.d. Print.

[83] Tekes. 2010. Automaatio liiketoimintaprosessien tukena. Tekesin katsaus 271/2010. Helsinki. 113 p.

[84] Palonen. 2014. Distributed Data Management of Automation System. Tampere. 12p

[85] Dretske, F. I.: Knowledge and the flow of information. Basil Blackwell Publisher, 1981.

[86] Heflin, J.: Towards the Semantic Web: Knowledge Representation in a Dynamic, Distributed Environment. Ph.D. Thesis, University of Mary land, College Park, 2001.

[87] Martensson, M. (2000). "A critical review of knowledge management as a management tool." Journal of Knowledge Management, 4, 204-216. Milton N. R.Knowledge Technologies. Milan: Polimetrica, 2008.

[88] KLIMEŠOVÁ, D. (2009). Data, Information and Knowledge Transformation.

[89] Bellinger G., 2004. Systems Thinking, The Knowledge Centered Organization, United States.

[90] Tampere University of Technology. 2014. *PLANTCockpit OS: Solution*. [ONLINE] Available at: http://www.tut.fi/plantcockpit-os/Solution.html. [Accessed 19 March 2017].

[91] McGilvray, D. 2008. Executing data quality projects ten steps to quality data and trusted information. Amsterdam, Morgan Kaufmann/Elsevier. 352 p.

[92] Redman, T.C. 2008. Data driven: Profiting from your most important business asset. Boston, Massachusetts, Harvard Business Press. 257 p.

[93] Norris, G., Hurley, J. R., Hartley, K. M., Dunleavy, J. R., & Balls, J. D. (2000). E-Business and ERP: Transforming the enterprise. New York: John Wiley & Sons, Inc

[94] Lieber, R. (1995). Here comes SAP. Fortune, 132(7), 122(3). Retrieved March 3, 2017.

[95] Minahan, T. (1998). Enterprise Resource Planning [ElectronicVersion].Purchasing, 125, 112. Retrieved March 3, 2017.

[96] Martinig& Associates. 2017. *Apache ServiceMix - Open Source SOA and Web Services*. [ONLINE] Available at: http://www.methodsandtools.com/tools/tools.php?servicemix. [Accessed 22 March 2017].

[97] Chary Eswarachary. 2012. *Integrating WebSphere Message Broker with Apache ActiveMQ*. [ONLINE] Available at: https://www.ibm.com/developerworks/websphere/library/techarticles/1211_eswarachary/1211_eswarachary.html. [Accessed 22 March 2017].

[98] Lambert, Douglas M., Martha C.Cooper and Janus D. Pagh, "Supply Chain Management: Implementation Issues and Research Opportunities," The International Journal of Logistics Management, Vol. 9, No.2 (1998), p. 1

[99] Fawcett, S., Wallin, D., Allred, C., Fawcett, A., &Magnan, G. (2011). Information technology as an enabler of supply chain collaboration: A dynamic - capabilities perspective. Journal of Supply Chain Management, 47(1), 38-59.

[100] Soh, C., Kien, S. S., &Tay-Yap, J. (2000). Cultural Fits and Misfits: Is ERP a Universal Solution? Communications of the ACM, 43(4), 47-51.

[101] Stevenson, W.J., 2007. "Operations Management", Ninth Edition, New York: McGraw-Hill. Subromanian S. (2008), "Commanding the internet era", Industrial Engineer: IE, Vol. 40 No. 10, pp. 44-8.

[102] Mabert, V.A., Soni A., Venkataramanan, M.A. 2003. "Enterprise Resource Planning: Managing the Implementation Process", European Journal of Operational Research 146, pp. 302-314